Clim. Past, 9, 1089–1110, 2013
www.clim-past.net/9/1089/2013/
doi:10.5194/cp-9-1089-2013

Climate of the Past
Open Access

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

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

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

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

Received: 31 May 2012 – Published in Clim. Past Discuss.: 22 June 2012
Revised: 28 February 2013 – Accepted: 16 April 2013 – Published: 16 May 2013

Abstract. Are simulations and reconstructions of past climate and its variability consistent with each other? We assess the consistency of simulations and reconstructions for the climate of the last millennium under the paradigm of a statistically indistinguishable ensemble. In this type of analysis, the null hypothesis is that reconstructions and simulations are statistically indistinguishable and, therefore, are exchangeable with each other. Ensemble consistency is assessed for Northern Hemisphere mean temperature, Central European mean temperature and for global temperature fields. Reconstructions available for these regions serve as verification data for a set of simulations of the climate of the last millennium performed at the Max Planck Institute for Meteorology.

Consistency is generally limited to some sub-domains and some sub-periods. Only the ensemble simulated and reconstructed annual Central European mean temperatures for the second half of the last millennium demonstrates unambiguous consistency. Furthermore, we cannot exclude consistency of an ensemble of reconstructions of Northern Hemisphere temperature with the simulation ensemble mean.

If we treat simulations and reconstructions as equitable hypotheses about past climate variability, the found general lack of their consistency weakens our confidence in inferences about past climate evolutions on the considered spatial and temporal scales. That is, 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 spatiotemporal climate variability in periods without instrumental coverage rely on two tools: (i) reconstructions from 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 (i.e. models) of varying complexity that produce discretely resolved spatiotemporal climate variables considered to represent a climate aggregation over regional spatial scales.

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 pseudo-observation, through the transfer function and through the assumption of a relatively stable "proxy"-climate relationship through time (e.g. Wilson et al., 2007; Bradley, 2011). Simulated climate estimates are uncertain due to the mathematical and numerical approximations of physical and biogeochemical processes (Randall et al., 2007). Additional uncertainty stems from the limited knowledge 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, for example, 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.

Thus we have no direct observational knowledge on the pre-industrial climate. Confidence in the inference on a past climate state requires reconciling models and reconstructions. As a first step towards this goal, we may apply methods from numerical weather forecast verification (see, e.g. Toth et al., 2003; Marzban et al., 2011; Persson, 2011) to evaluate the consistency of an ensemble of estimates with relevant verification data. These methods are less subjective than by-eye evaluations. Practically, we select a verification data target from the available reconstructions to verify an available ensemble of simulations, and vice versa. For a specific task at hand, the analysis identifies whether the ensemble and the verification target are *consistent* realisations of the unknown past climatology or of the unknown past climate evolution. We consider ensemble consistency as used in the field of weather-forecast verification (e.g. Marzban et al., 2011). Such consistency has to consider probabilistic and climatological properties (see below for a more detailed explanation of these two terms in the current context). Reconstructions and simulations are therefore treated as different but equitable hypotheses, and their consistency is assessed within the framework of a statistically indistinguishable ensemble (Toth et al., 2003). The concept of indistinguishability or exchangeability bases on the assumption that the true climate system or the target system is sampled from a distribution of model systems (compare, e.g. Toth et al., 2003; Annan and Hargreaves, 2010; Sanderson and Knutti, 2012). The target and the ensemble are indistinguishable with respect to their statistics if they are sampled from the same (or at least similar) distributions. Note, to be consistent does not imply to be identical (see for example Annan et al., 2011).

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. Annan and Hargreaves (2010) and Hargreaves et al. (2011) discuss, respectively, the 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 ensembles of reconstructed and simulated estimates of 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 the temporal evolutions over the last millennium of the COSMOS-Mill simulation ensemble with reconstructions for Central European mean temperature (Dobrovolný et al., 2010) and a temperature field reconstruction (Mann et al., 2009). Probabilistic reconstruction-simulation consistency is assessed using rank histograms (e.g. Anderson, 1996) and the decomposition of the χ^2 goodness-of-fit test statistic (Jolliffe and Primo, 2008). The climatological component of ensemble consistency is evaluated by presenting residual quantile-quantile plots (Marzban et al., 2011; Wilks, 2011). The methods are discussed in Sect. 2. Section 3 presents results concerning the consistency of reconstructions and simulations. Appendix B discusses the robustness of the approach.

2 Methods and data

This study details case studies for the validation of ensembles of paleoclimate estimates from simulations and reconstructions. We assess whether the ensembles of interest can be considered to be consistent with a relevant verification data target. We detail herein criteria for rejecting such consistency and evaluate the ensembles with respect to these criteria. This section introduces the methods and the criteria.

An ensemble of (climate) estimates can be validated against a suitable verification data target either by considering individually the accuracy of each ensemble member or by evaluating the consistency of the full ensemble (e.g. Marzban et al., 2011). The latter approach may follow the methods applied in the verification of numerical weather forecast ensembles. These are based on the concept of statistical indistinguishability of the ensemble, which is interpreted probabilistically. We assume that the verification target and the simulations are samples from a larger distribution (compare, e.g. Toth et al., 2003; Annan and Hargreaves, 2010; Persson, 2011; Sanderson and Knutti, 2012) so that their statistics are exchangeable.

Practically, exchangeability – or, analogously, 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. For a consistent ensemble, the verification target and ensemble estimated (e.g. forecasted) frequencies agree (Murphy, 1973). Thus an ensemble is probabilistically consistent if we cannot reject the hypothesis that the frequencies agree. In forecast ensemble verifications, an ensemble is called reliable, if we cannot reject this hypothesis according to appropriate tests. Since we deal with highly uncertain data, we do not use the term "reliability" here but only refer to consistency. The assessment of ensemble consistency provides a necessary condition for our evaluation of ensemble accuracy in paleoclimate-studies (following Annan and Hargreaves, 2010) under large uncertainties and due to the lack of an observed target.

Besides the probabilistic ensemble consistency, and following Marzban et al. (2011) and Johnson and Bowler (2009), we have to additionally consider the climatological consistency of the ensemble members. That is, we need not only to evaluate whether within-ensemble frequencies are consistent with those of the verification data, but also whether the variance of the ensemble members' climatologies agree with the verification climatology.

2.1 Methods

2.1.1 Evaluation of probabilistic consistency

Probabilistic consistency is commonly evaluated by ranking the verification target data against the ensemble data (Anderson, 1996; Jolliffe and Primo, 2008; Annan and Hargreaves, 2010; Marzban et al., 2011; Hargreaves et al., 2011). Target data and ensemble data are sorted by value and the calculated ranks counted and plotted as a rank histogram (Anderson, 1996). If we expect equiprobable outcomes for an ideal, indistinguishable ensemble, the ranking should result in a uniform, flat histogram.

We can test the goodness-of-fit of a rank histogram relative to the flat expectation, i.e. with respect to the null hypothesis of a uniform outcome. An ensemble is probabilistically consistent if we fail to reject the hypothesis of a uniform histogram. One suitable test is the χ^2 test (e.g. Jolliffe and Primo, 2008). Jolliffe and Primo (2008) detail how we can decompose the test statistic enabling tests for individual deviations from flatness which are due to different statistics of the distributions. Please see Jolliffe and Primo (2008) for a comprehensive delineation. Appendix A presents more details on the test and discusses the chosen approach.

Besides the possibility to test for deviations from a uniform outcome, the rank histograms already visually assist in identifying discrepancies between the ensemble data and the verification data. An apparent dome-shaped histogram indicates that the ensemble data is sampling from a distribution which is wider than the verification data distribution. A u-shape signals an ensemble distribution which is narrower than the verification data distribution. The spread of the ensemble is, respectively, overly wide or overly narrow. That is, the ensemble data differs in its variance from the verification data. We refer to too wide distributions as being over-dispersive and to too narrow distributions as under-dispersive. If the ensemble is biased to large values, the rank counts display a negative trend. If it is biased to low values, the trend is positive in the rank counts. Consequently, if the ensemble data has a negative bias, the verification data will cluster in the high classes of the histogram and vice versa. Jolliffe and Primo (2008) give details on other possible deviations, but we focus here only on biases and differences in the spread of the ensemble data.

In summary, rank histograms are a tool to disclose whether a probabilistically interpreted ensemble and its verification data represent different climates. They provide a means for evaluating the consistency of the joint distribution for the ensemble and verification data (see Wilks, 2011).

The ranking further allows mapping the ranks of the verification data and thereby helps in validating gridded spatial data. That is, the position of the verification data within the ensemble can be visualized in maps (see Sects. 3.2.1 and 3.2.2). The rank of the verification data is plotted at each grid-point for individual time steps or for climatological averages. Local low rank counts of the target indicate that the ensemble is biased high at the grid-point, and high ranks imply a low bias.

2.1.2 Evaluation of climatological consistency

Following Marzban et al. (2011, see also Wilks, 2011), we use residual quantile-quantile plots (r-q-q plots) to study the climatological consistency of the distributions for the ensemble with the target. Common quantile-quantile plots assess the quantiles of a distribution against a reference. For example, the quantile estimates of a hindcast simulation are plotted on the y-axis against the observed quantiles on the x-axis. Residual quantile-quantile plots only differ from this common approach by plotting the differences between the simulated distribution quantiles and the chosen verification data quantiles on the y-axis. Thereby they emphasise the deviations between the simulated and the verification quantile distributions.

This visualisation allows assessing whether the climatological distribution of an estimate of interest is similar to the distribution of the target. Thus we are able to identify whether the empirical quantiles for each individual ensemble member agree with the verification data sample. Plotting the residuals eases the interpretation since ideal agreement between estimated and verification quantiles leads to vanishing residuals, i.e. a horizontal line crossing the y-axis at zero. Thus disagreements are also easily identified. Differences in the tails of the distributions, their skewness or their means are of particular interest among the possible deviations.

Biases of the estimated distributions lead to horizontal displacements from the expectation of vanishing residual quantiles in the residual quantile-quantile plot since the mean of the estimated distribution differs from the verification reference. The residuals show a positive slope if the estimated climatological distribution is wider than the verification climatology distribution, and a negative slope if it is narrower (Marzban et al., 2011). This reflects differences in estimated and target climatological variances. Thus if the climatological variance of the estimate is larger than that of the target, the ensemble systematically overestimates the distance between the mean and the quantile locations. This results in a positive slope in the residual quantile. Smaller climatological variance results in a negative slope since the quantiles are closer to the mean. Marzban et al. (2011) give more details on the interpretation of the pattern of residual quantiles. We refer to differences in the width as over- and under-dispersion for too wide and too narrow distributions, respectively.

In summary, quantiles or residual quantiles account for differences in the climatologies of the ensemble members and thereby complement the analysis of probabilistic ensemble consistency. In climate studies, they are especially useful to highlight differences in the resolved values close to the tails. While the rank histograms consider the joint distributions (see above), the residual quantiles highlight differences of the climatologies of individual simulations with the verification data.

2.2 Data

We apply the described methods to the following data sets. The simulations are the COSMOS-Mill model data for the last millennium performed with the Max Planck Institute Earth System Model (MPI-ESM). This version of MPI-ESM is 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. Furthermore, there are five full-forcing simulations with weak solar forcing and three full-forcing simulations with strong solar forcing. The full ensemble has eleven members. We assume that our estimates of the forcing series are highly uncertain and that this uncertainty propagates to our knowledge of their influence on the pre-industrial climate. Therefore, we include the single forcing simulations as valid hypotheses about the pre-industrial climate trajectory. We denote the full ensemble by SIM. WSIM and SSIM refer to the full-forcing simulations with weak and strong solar forcing, respectively (i.e. the weak and strong ensembles). Additionally, we take advantage of the 3100 yr control run describing an unperturbed climate.

The reconstructions are all for annual mean temperature. We use the regional mean time series for Central Europe by Dobrovolný et al. (2010), the ensemble of Northern Hemisphere means by Frank et al. (2010) and the global field reconstruction by Mann et al. (2009). All series have an annual resolution, but some are temporally smoothed (e.g. Mann et al., 2009).

The Frank et al. (2010) data is the only available ensemble of reconstructions. Frank et al. (2010) recalibrate a number of previous reconstructions to various periods of instrumental observations, thereby obtaining an ensemble of 521 recalibrated reconstruction series (see Frank et al., 2010, for discussions on the ensemble construction). The ensemble bases on 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). The original reconstructions end at different dates, that is, their last available annual data differ. We refer to the full 521 member ensemble as FRA. The choice of the calibration window strongly influences the variability of the reconstructions which is going to influence subsequent analyses. The 1920–1960 period likely presents the most reliable observational data if we want to use all nine reconstructions. Therefore, in the following, we use the subensemble re-calibrated to the period 1920 to 1960 and refer to it as FRS.

We interpolate the spatial field data on a 5° × 5° grid. Our general interest is in the consistency of paleoclimate reconstructions and simulations for the last millennium. Therefore we take anomalies with respect to the common period of reconstructions and simulations but exclude the period of overlap with the modern observations. European anomalies are for the period 1500 to 1854 relative to the mean from 1500 to 1849. For the Northern Hemisphere data, we compute anomalies for the period 1000 to 1849 and relative to the mean for the same period. The decadally smoothed global fields for the years 805 to 1845 are centred relative to the mean for the period 800 to 1849. We further consider four sub-periods in the analyses of the global field data. These consist of 250 non-overlapping records from the full period 805 to 1845. The first three periods cover the first 750 records, and the last period covers the last 250 records.

The rank histogram approach (see Sect. 2.1) assumes that the validation data sets include errors (Anderson, 1996) that have to be included in the ensemble data. If the reconstructions are reported with an uncertainty estimate, this is used to inflate the simulated data.

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). No uncertainty estimate is given for the global field data. However, Mann et al. (2009) provide standard errors for their unscreened Northern Hemisphere mean temperature series. We assume that the largest standard error reported for this data is a reasonable guess for an uncertainty estimate for the field data as well. Thus we choose to inflate the ensemble at each grid-point by a random uncertainty estimate drawn from a Gaussian distribution with standard deviation equal to this standard error (i.e. $\sigma = 0.1729$).

The SIM and the FRS data are ensembles. Thus we can randomly sample an "observational" uncertainty for their ensemble means from a distribution with zero mean and standard deviation equal to the ensemble standard deviation at each point. For the analyses relative to an ensemble mean, we additionally use additive internal variability estimates for the target data (see Sect. 2.3 for details).

As mentioned above, the way we construct the FRS ensemble influences the ensemble spread and thus the results. We account for this sensitivity by basing the uncertainties for the ensemble-mean reconstruction on the full FRA-ensemble spread.

2.3 Discussion of the chosen approach

Already the first applications of the rank histogram advised caution in its interpretation (e.g. Anderson, 1996) not least because of the uncertainties in the verification data. More recently Hamill (2001), Marzban et al. (2011) and others discussed the influence of, for example, the underlying distributions or temporal correlations on the results; see also Wilks (2011) and the references in these publications. Marzban et al. (2011) further discuss the influence of intra-ensemble correlations and correlations between the ensemble and the verification target on the rank histogram.

One of the most important limitations of the rank histogram approach is that uniformity in rank histograms may result from opposite biases or opposite deviations in spread in different periods which may cancel out (Hamill, 2001). Furthermore, temporal correlations in the data can result in premature rejection of consistency (Marzban et al., 2011). We use bootstrapped estimates and analyse different sub-periods at individual grid points to address these problems.

Rank histograms may be misleading since they are affected by the amount of correlations between verification and ensemble or within the ensemble and the differences between both (Marzban et al., 2011). We assume that these caveats increase the general uncertainty in the comparison between simulations and reconstructions of past climate states and variability.

We may evaluate the consistency of simulations and reconstructions on three levels of resolution: area-averaged time series, gridded spatio(-temporal) data, or individual grid points of gridded data sets. Obviously, results may differ between these levels. Further, it is not easy to know a priori whether we should expect larger consistency at one of the resolutions. Note that 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 is consistent with the covariability in the verification data.

We assume that the data sets represent inter-annual variations of a temperature index. This is not necessarily valid. If the target is an ensemble mean, it displays reduced variability compared to the ensemble members especially on the inter-annual time scale. Thus using an ensemble mean as verification target impacts the ensemble consistency. We argue that the inherent uncertainty of the target may compensate for this reduced variability caused by ensemble averaging. Assuming that reconstruction and simulation ensemble estimates include the same externally forced variability, the target ensemble mean should essentially recover the forced signal within the propagated uncertainties. Furthermore, the probabilistic ensemble estimates should reliably represent the target distribution if the ensemble includes the target uncertainty.

Nonetheless, in the following we pursue an alternative approach to compensate for the reduced variability of an ensemble-mean target and add an estimate of the internal variability to the ensemble-mean estimate. In assessing the consistency of the SIM ensemble, we first compute the residual deviations of the full FRA-reconstruction ensemble from its ensemble mean. Then, we fit autoregressive-moving-average models to the residuals. Thereby we obtain 521 possible fits. We produce 10 random representations of the process for each fit. If we add these 5210 estimates of residual internal variability to the ensemble mean, we obtain a set of targets. For the assessment of the FRS ensemble, we add one section of the MPI-ESM control run (Jungclaus et al., 2010) to the SIM-ensemble mean. Since we further account for the sampling variability, using one segment is robust enough for evaluating the unforced internal variability of the simulations.

Finally, FRA- and FRS-ensemble members are to some extent time-filtered. They exhibit by construction reduced variability on inter-annual time-scales (compare, e.g. Franke et al., 2013). As the filter-properties differ, we do not account for this filtering. On the other hand, we use decadal moving means for the SIM data to compensate for the decadal smoothing of the global field data (Mann et al., 2009).

Appendix B supplements our assessment of mutual simulation-reconstruction consistency by presenting evaluations of the self-consistency for both the control run and the FRS ensemble. The first allows assessing the self-consistency of the unperturbed simulated climate and the sensitivity of our tests. The second allows evaluating the large uncertainty in our reconstructed estimates and of the available targets for evaluating the simulations.

3 Results

We first evaluate the consistency of the SIM ensemble relative to two reconstruction targets: the Central European temperature data by Dobrovolný et al. (2010) and the ensemble mean of the Northern Hemisphere temperature FRA ensemble. In a reverse analysis, we then test whether the FRS ensemble is consistent with the ensemble mean of the SIM-ensemble. SIM, WSIM and SSIM are further analysed for their consistency with individual members of the FRS-data. Finally, we assess the ensemble consistency of the SIM-ensemble with the global field reconstruction by Mann et al. (2009).

3.1 Area-averaged time series

Figures 1 to 3 provide a first impression of the ensemble data sets and the respective verification target series. We display the target time series and their variability together with the range of the ensembles.

The European data for the SIM ensemble and its reconstructed verification target cover a similar range and show similar variability (Fig. 1). Note that the SSIM and WSIM ensemble means exhibit a reduced variability compared to the full ensemble range of variability (Fig. 1b).

On the other hand, the northern hemispheric SIM ensemble data varies more than its verification target which is the FRA ensemble mean. The verification target also displays a different temporal evolution (Fig. 2). Similar differences occur when comparing the FRA ensemble and its verification target, i.e. the SIM ensemble mean (Fig. 3). However, here the verification data variability is in the range of the ensemble variability.

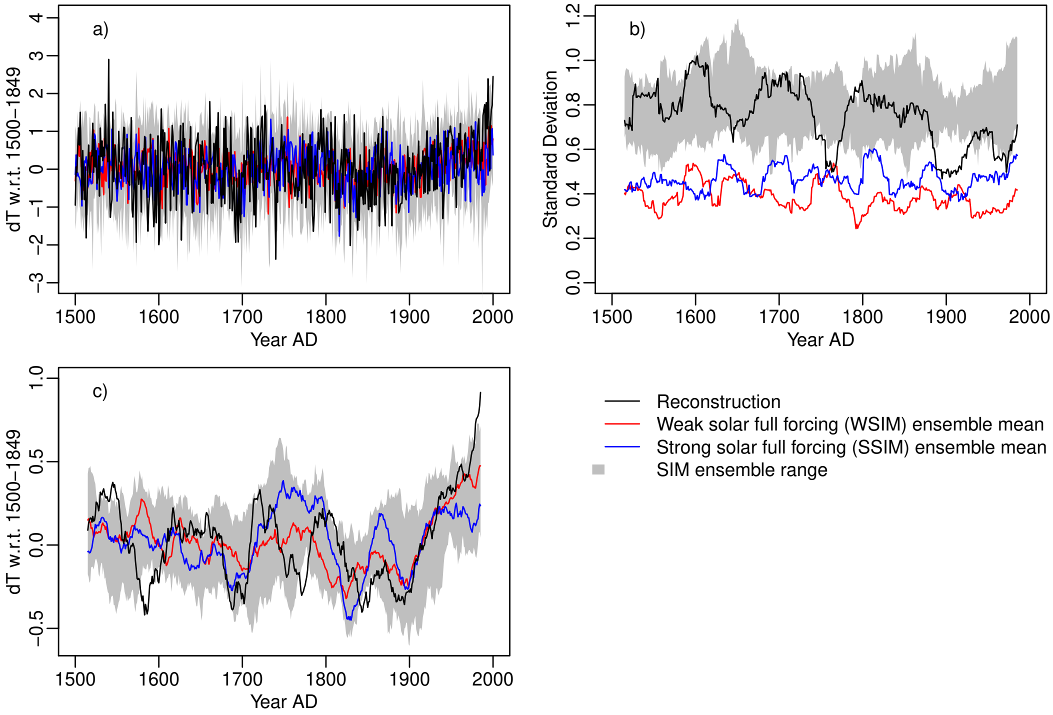

Fig. 1. (a) Time series, **(b)** moving 31 yr standard deviations and **(c)** moving 31 yr means for the Central European annual temperature data. Black is the target data and transparent light grey shading is the range of the ensemble. Red (blue) lines are for the WSIM (SSIM) full-forcing simulation ensemble means.

Including the estimates of internal variability increases the range of possible temporal evolutions of the reconstructed verification targets for the SIM ensemble (Fig. 2). In contrast, the verification target for the FRS ensemble does not change too much if we include an estimate of internal variability (Fig. 3). Nevertheless, we see a pronounced increase in the variability of the SIM ensemble mean. Sections 3.1.2 and 4 discuss the influence of the resolved variability on the results.

Figures 4 and 5 illustrate the analyses of, respectively, the probabilistic and the climatological consistency for these three ensemble-data sets. Both figures account for the uncertainty in the verification target as described in Sect. 2.2. Uncertainty estimates are the reported standard errors for the Central European temperature target (Dobrovolný et al., 2010) and the spread of the mutual ensembles for the Northern Hemisphere data targets. If we neglect these "observational" uncertainties in the verification data the conclusions change for the hemispheric data (not discussed).

3.1.1 Ensemble consistency of area-averaged estimates

The rank counts plotted in Fig. 4a indicate that we cannot reject consistency of the SIM ensemble data for the annual mean Central European temperature with the reconstruction verification target. Nevertheless, the test statistics for a deviation in spread are significant, implying a lack of consistency. Interestingly, the bootstrapped rank count intervals enclose the possibility of a uniform rank count, i.e. of consistency. The contrast between bootstrap and goodness-of-fit test possibly highlights the problem of sampling variability.

The results differ for the assessment of probabilistic consistency of the hemispheric estimates depending on whether or not we account for internal variability in the ensemble-mean target data (Fig. 4b,c). We first consider the case where the assessment does not include the estimates of internal variability described in Sect. 2.3. Then, for the evaluation of the SIM ensemble, the dome-shaped histogram in Fig. 4b shows that the verification target occupies too often the central ranks, i.e. the SIM ensemble is significantly over-dispersive. The bootstrapped intervals confirm this (cyan overlay in Fig. 4b). Similarly the FRS ensemble is too wide relative to the Northern Hemisphere target of the simulation-ensemble mean if we do not account for the reduced internal variability (cyan overlay in Fig. 4c).

However, results are ambiguous if we include the estimates of internal variability. SIM appears to be consistent with some of the targets, but the summarising statistics for the assessments against all targets still emphasise an apparent over-dispersive relation. For example, the 90 % envelope is incompatible with a uniform histogram, it indicates a lack of probabilistic consistency (dark grey in Fig. 4b), but we cannot reject consistency according to the 99 % envelope (light grey in Fig. 4b).

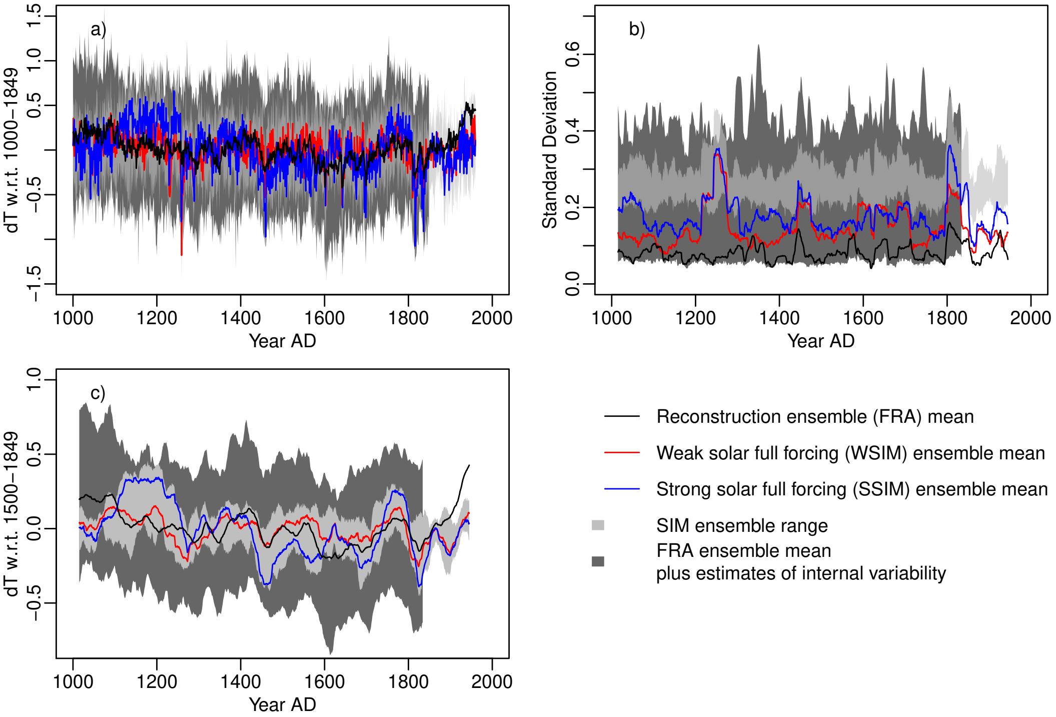

Fig. 2. **(a)** Time series, **(b)** moving 31 yr standard deviations and **(c)** moving 31 yr means for the SIM Northern Hemisphere temperature data against the reconstructed target. Black is the verification data target and transparent light grey shading is the range of the SIM ensemble. Red (blue) lines are for the WSIM (SSIM) full-forcing simulation ensemble means. Dark grey shading is the range of the reconstruction ensemble-mean target with added internal variability estimates. Here and in Fig. 3, the data including the estimate for internal variability is only shown for the period of analysis from the start of the millennium to the mid-19th century.

On the other hand, FRS is consistent with the ensemble-mean simulation target when we include the estimate of internal variability for the simulation. That is, deviations from a flat histogram are negligible for the continuous black line in Fig. 4c and the test statistics are also not significant. The bootstrapped ranks (grey shading in Fig. 4c) further highlight the good probabilistic agreement under the made assumptions. The reconstructions by Hegerl et al. (2007), Mann and Jones (2003) and Mann et al. (2008) are filtered estimates, but the conclusions do not change if we include an arbitrarily chosen estimate of internal variability in these three series.

The climatological quantiles support the probabilistic assessment. They agree rather well between the SIM ensemble members and the Central European temperature target data. The residual quantiles align more or less close to zero in Fig. 5a. Some simulations appear to underestimate very warm anomalies and overestimate very cold anomalies. A slight positive slope occurs in the residual quantiles but we conclude that this over-dispersive tendency is not significant since the bootstrapped intervals still include the zero line.

The climatological deviations between the quantiles for the Northern Hemisphere temperature in SIM and the target are larger than for the Central European data. The SIM members give positively sloped residual quantiles, i.e. overly wide climatological distributions, if we do not account for the reduced internal variability in the relevant target (the FRA ensemble mean) (grey overlay in Fig. 5b). Similarly, FRS ensemble members generally overestimate at least the positive anomaly quantiles relative to the target (the SIM ensemble mean) if we exclude the internal variability estimate (transparent grey in Fig. 5c).

The results change if we include the internal variability estimates. Figure 5b displays residual quantiles for the SIM ensemble relative to the targets including estimates of internal variability (see Sect. 2.3). It is apparent that there are consistent, near-vanishing residuals but also negatively sloped under-dispersive or positively sloped over-dispersive cases. Quantiles in the tails appear to commonly agree between the SIM members while the variability closer to the mean of the distribution is overestimated. That is, residuals are small in the tails but display a positive slope for more central quantiles. There are also cases for which SIM ensemble members are more variable close to the mean but the tails are lighter compared to the target. The overestimation of the variability appears to be largest for the SSIM sub-ensemble simulations (i.e. simulations which use a strong solar forcing). From our point of view, the multitude of possible deviations requires to conditionally reject the hypothesis of climatological consistency. This is mainly due to the notable overestimation of variability. This is in line with the probabilistic assessment

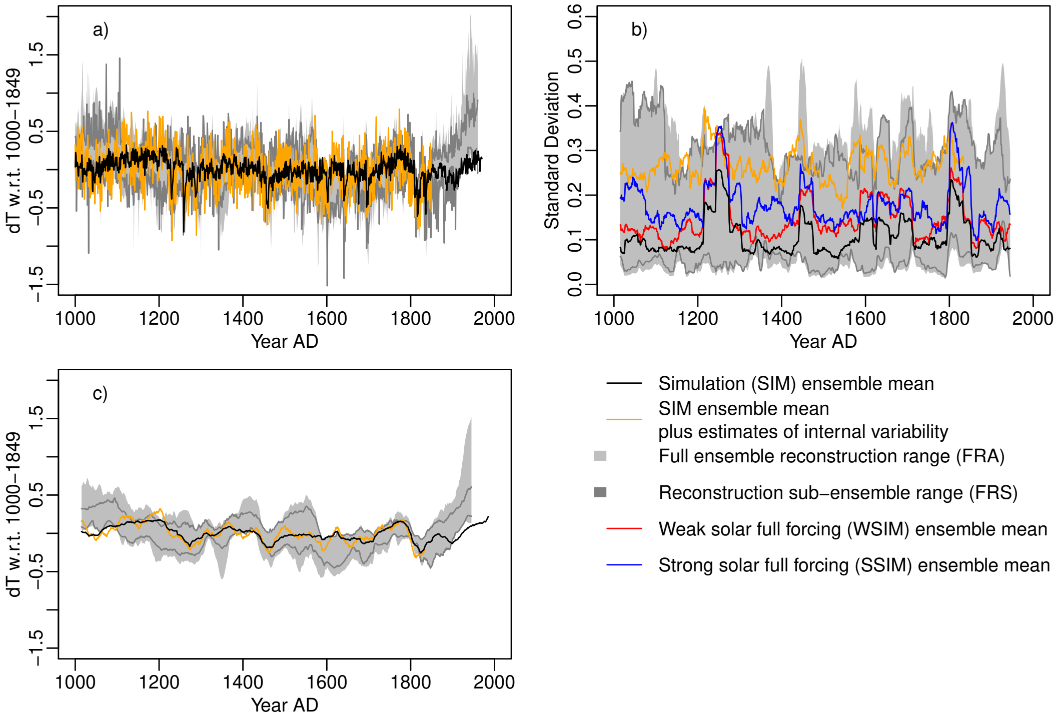

Fig. 3. (a) Time series, **(b)** moving 31 yr standard deviations and **(c)** moving 31 yr means for the FRA Northern Hemisphere temperature reconstruction ensemble against the simulated target. Black is the verification data target and transparent light grey shading is the range of the FRA ensemble. Dark grey lines mark the range of the FRS reconstruction sub-ensemble recalibrated to the period 1920–1960. The orange line is the estimate of the target with added internal variability estimate. In **(b)** red (blue) lines are for the weak (strong) solar full-forcing simulation ensemble means.

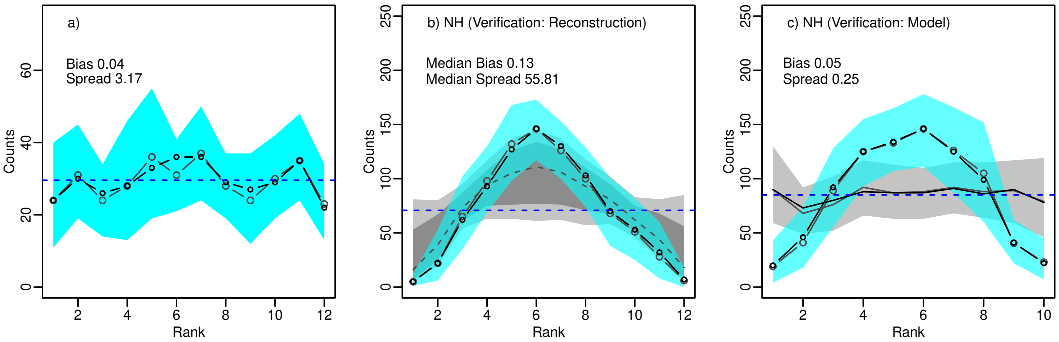

Fig. 4. Evaluation of probabilistic consistency: rank histogram counts (black line with points) for temperature data: **(a)** SIM against Central European annual temperature target, **(b)** SIM against Northern Hemisphere temperature target, **(c)** FRS against Northern Hemisphere temperature target. 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. Cyan shading (grey lines with points) is 0.5 % and 99.5 % (50 %) quantiles for block-bootstrapped rank histograms (2000 replicates, block length of 50 yr) relative to raw targets. 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 relative to the ensemble-mean target 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.

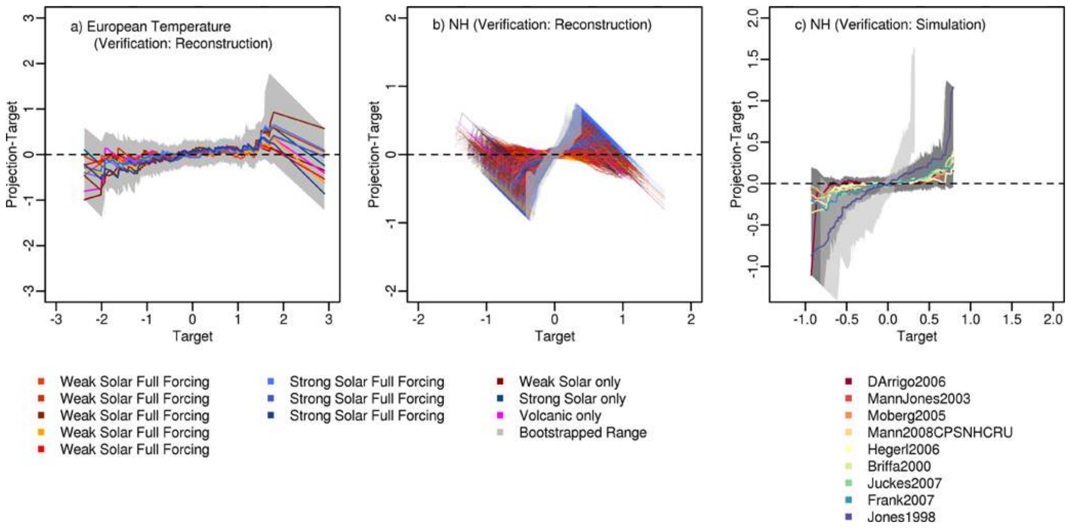

Fig. 5. Evaluation of climatological consistency: residual quantile-quantile plots for temperature data: **(a)** SIM members against Central European annual temperature target, **(b)** SIM members against Northern Hemisphere temperature target, **(c)** FRS against Northern Hemisphere temperature target. Panels account for the uncertainties in the 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.

(see above, Fig. 4b) where we also find a generally over-dispersive character of the SIM ensemble.

For the FRS ensemble, we generally find good agreement between the quantiles of the ensemble members and the simulation target if we include an estimate of internal variability in the target (Fig. 5c). For most members, large residuals occur only in the tails of the distribution. The bootstrapped intervals emphasise this general consistency by including the zero line of vanishing residuals. The deviations in the tails are most pronounced for large negative anomalies in the reconstruction by D'Arrigo et al. (2006). An exception to this general description is the data set by Jones et al. (1998). For this reconstruction a strong positive slope in the residuals indicates a strong over-dispersive character. Much of the over-dispersion comes from the large associated uncertainties.

The next paragraphs complement the above results by shortly looking at some sub-divisions of the considered SIM and FRS ensemble. Since the SIM ensemble encapsulates the SSIM and WSIM ensembles, we shortly discuss the consistency of these two sub-ensembles. We consider the uncertainty and, for the hemispheric data, also include internal variability estimates. For the sake of brevity, we just report the results. Generally, results for the two sub-ensembles agree well with those found for the full SIM ensemble relative to the Central European temperature. However, both, SSIM and WSIM, display specific behaviours. WSIM is unambiguously probabilistically consistent with the European reconstructions, but SSIM is slightly too wide. SSIM deviations in the spread are significant according to the goodness-of-fit test. However, the bootstrapped intervals suggest that this may be due to sampling variability. Results for the climatological consistency are similar for SSIM and WSIM as seen in the SIM assessment in Fig. 5a.

With respect to the Northern Hemisphere mean target, the WSIM ensemble is probabilistically too wide while we are only able to make ambiguous statements for SSIM. Since the SSIM ensemble has only three members, we have anyway to be careful when interpreting the results. The single deviation test for spread suggests significant over-dispersion. However, the bootstrapped rank intervals do not allow rejecting consistency since they safely include the possibility of a flat histogram. The residual quantiles display a wide range of possible deviations for SSIM (compare Fig. 5b).

For the reversed verification, the single deviation tests indicate significant spread deviations of the FRS reconstruction ensemble. It is slightly too narrow if the target is the WSIM ensemble mean, but strongly too narrow if the target is the SSIM ensemble mean. However, the bootstrapped intervals again allow for consistency relative to both SSIM and WSIM. Climatologically, most FRS ensemble members are consistent with both targets but again the results are distinct for the reconstruction by Jones et al. (1998). That is, for all FRS members the climatological deviations relative to the SSIM and WSIM ensemble-mean targets are similar to those relative to the SIM ensemble mean (compare Fig. 5c), but the residuals are larger when evaluated against the SIM ensemble-mean target.

While we may interpret the ensemble-mean reconstruction as "best available" target for verifying the SIM ensemble, we should also consider the consistency of the simulation ensemble relative to individual reconstructions. We shortly present results on the assessment of SIM, SSIM and WSIM relative to the FRS ensemble members as targets. Here, we include uncertainty estimates. Furthermore, we add an arbitrary member of the ensemble of estimates of internal variability estimates (see Sect. 2.3) to the three filtered reconstructions by Hegerl et al. (2007), Mann and Jones (2003) and Mann et al. (2008). Figure 6 presents the χ^2 values for the tests.

Obviously, the SIM ensemble lacks probabilistic consistency with all reconstructions according to the χ^2 test (considering our assumptions on internal variability and uncertainty). The bootstrapped intervals confirm this (not shown). The full test gives no significant results relative to the reconstruction by Moberg et al. (2005). Climatological quantiles confirm the probabilistic findings (not shown).

WSIM appears to be probabilistically consistent with the Moberg et al. (2005) reconstruction. The bootstrapped intervals suggest that the ensemble is not inconsistent with the data by Mann et al. (2008) under the made assumptions (not shown). Residual quantiles are large except relative to the reconstruction by Moberg et al. (2005).

The three-member SSIM ensemble is a special case. Bootstrapped intervals do not allow to reject probabilistic consistency for any of the nine reconstructions under the assumptions on internal variability and uncertainty. Test statistics in Fig. 6c 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).

In summary, verification of the SIM ensemble suggests that it is likely too wide relative to the ensemble mean of the Northern Hemisphere mean temperature reconstructions. Discrepancies arise not only probabilistically but also in the climatologies. The climatological results may depend on the representation of the internal variability in the verification target. The FRS ensemble for Northern Hemisphere temperature, on the other hand, appears to be consistent with its target (the SIM ensemble mean) if we account for uncertainties and internal variability. Nevertheless, most ensemble members display climatological deviations in the tails of the distribution. In the end, the large uncertainties in the ensembles and in the verification targets generally prohibit rejecting consistency for the Northern Hemisphere estimates. The results are more encouraging for the considered regional temperature estimate. The SIM estimates for the Central European temperature appear to be unambiguously consistent with the respective target.

3.1.2 Addressing lack of consistency of area-averaged estimates

Returning to Figs. 1 to 3, the following notes are worth repeating. First, while the SIM ensemble covers a similar range of temperature values as the Central European reconstruction target and while their variability is also similar, the low-frequency variability differs notably between the ensemble and the target (Fig. 1b). Secondly, differences between SIM and the northern hemispheric target are prominent and also between FRS and its target (Figs. 2 and 3). Furthermore, with respect to the hemispheric data, the range of the reconstructed targets is relatively wide compared to the SIM ensemble spread if we account for the reduced internal variability in the original FRA ensemble-mean target. The moving standard deviations emphasise the disagreement in variability (Fig. 2b). For the FRS ensemble, on the other hand, including an estimate of internal variability does not unduly change the respective target (Fig. 3a). However, the variability of the target increases notably (Fig. 3b).

Thus ensemble data can be statistically indistinguishable from a verification target although their trajectories evolve notably different over much of the considered time-span. This is seen for the Central European temperature data (Figs. 1c, 4a, 5a) in both the probabilistic and the climatological assessment. That is, the strong differences in the 18th century (or similarly the late 1500s) are consistent with our knowledge about internal and externally forced climate variability for the continent under the uncertainties associated with reconstructions, climate simulations and the forcing reconstructions.

This obviously does not hold for the hemispheric data of the SIM ensemble for which the probabilistic and climatological evaluations reveal disagreements with the target. The time series in Fig. 2 clarify that part of the over-dispersive character of the hemispheric SIM data may relate to (i) biases in the periods 1000 to 1300 and 1500 to 1650, when reconstructions and simulations evolve to some extent in opposite way, and to (ii) less warming in the target in the 18th century. The same biases act in opposite directions in the evaluation of the hemispheric data of the FRS ensemble. However, here the biases are not large enough to allow rejecting consistency. Indeed, they compensate over the full period.

Overall, our analyses depend on an appropriate representation of internal variability, which may be as large as the forced signal amplitude. Our approaches to include internal variability differ for the analysis of the SIM and FRS ensembles. Results for the hemispheric SIM and FRS ensembles describe different aspects of our uncertain knowledge even after accounting for the reduced variability in their respective targets. The spread of the reconstruction ensemble relates to different methodologies and different climate proxies, but the simulation intra-ensemble variability represents the differences in the considered forcing estimates and the different initial conditions of the ensemble. For the simulations, the spread also depends on the formulation of the numerical code. The latter is a smaller issue in the present study but becomes important for multi-model ensembles. Thus for an ensemble-mean simulation as target, our internal variability estimate describes one unperturbed climate trajectory under similar constraints. The internal variability adjustments

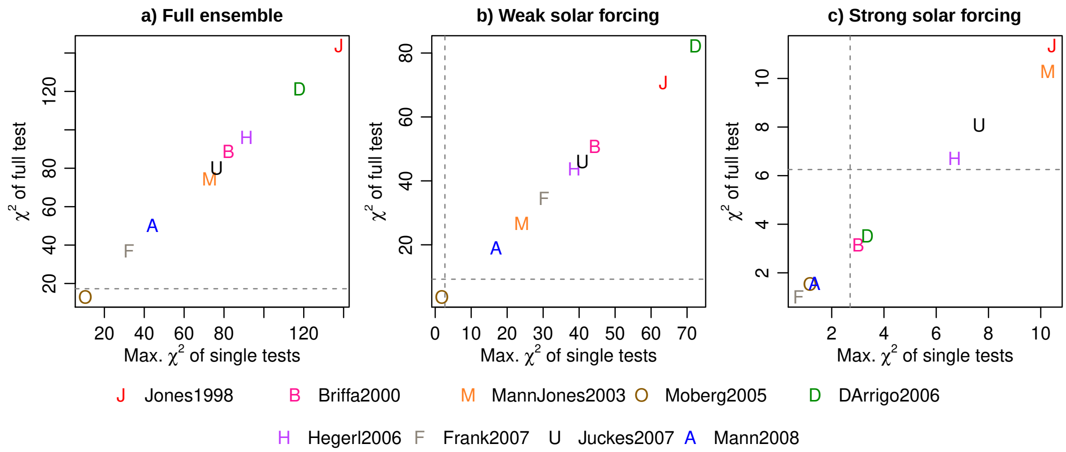

Fig. 6. Assessing SIM, WSIM and SSIM ensembles against individual reconstructions of Northern Hemisphere temperature (members of FRS ensemble). Uncertainties are considered, and internal variability estimates are included in the data by Hegerl et al. (2007), Mann and Jones (2003) and Mann et al. (2008) to account for the temporal filtering of the individual reconstructions. **(a)** SIM: χ^2 statistics for the full test against the maximum of the decomposed χ^2 statistics obtained for the tests for bias and spread deviations. **(b)** as in **(a)** but for the five-member WSIM ensemble. **(c)** as in **(a)** but for three-member SSIM. Vertical and horizontal grey lines mark those χ^2 statistics for which left p values are larger than 0.9 for the distributional degrees of freedom.

for an ensemble-mean reconstruction as target still represent the different methodologies and the different types of proxy data although the estimates are generated as stochastic processes.

Our presented analyses deal with data sets which either have similar variability (the Central European data) or for which we have to account for reductions in internal variability since we employ ensemble-mean targets. However, strong discrepancies occur also between SIM and reconstructions which are resolved at inter-annual time-scales (not shown).

3.2 Spatial fields

Figure 7 gives some information on the global temperature data for an arbitrarily chosen sub-period (1390s to 1690s) as they are depicted by reconstructions (Fig. 7a) and simulations (ensemble mean in Fig. 7b, compare also Fig. 7 of Fernández-Donado et al., 2013). Comparison of Fig. 7a and b highlights how strongly mean anomalies of the SIM ensemble may disagree with the target pattern for this specific period.

While the ensemble mean, of course, smoothes out the patterns found in individual runs, it is noteworthy, that not only the SIM ensemble but also a multi-model ensemble (Fernández-Donado et al., 2013, compare their Fig. 7) capture basically none of the reconstructed features for this period. This potentially highlights the limited value of such simple comparisons. The most prominent mismatch between the ensemble and the target is found in the tropical Pacific (compare Fig. 7a, b, d). 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 warm mean anomaly of the target and the diverse but generally much weaker mean anomalies in the SIM ensemble. Simulations incorporating a strong solar forcing even display negative anomalies (not shown). Such a La Niña-like response not only conflicts with the target, but also contrasts with the findings during solar forcing minima by Meehl et al. (2009) and Emile-Geay et al. (2007); see also the discussions by Misios and Schmidt (2012) on the relationship between solar insolation maxima and tropical Pacific sea surface temperatures.

In the following, we evaluate the consistency of the SIM ensemble relative to the decadally smoothed global temperature fields. We repeat that deviations from a uniform rank histogram count may be due to biases or differing spread in particular periods, while the ensemble is consistent with the target in other periods. Discrepancies can easily be identified when analysing single time series but assessments of consistency are not easily visualised at the grid-point level of spatial fields. We use different time periods to account for possible changes in deviations over time. We employ sub-periods of non-overlapping 250 records in the range from 805 to 1845. The first three periods cover the first 750 records of the full data (about 805 to 1055, 1055 to 1305, 1305 to 1555) while 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.

No uncertainty estimate is given for the global field target (reconstruction by Mann et al., 2009). We consider the largest standard error of the unscreened Northern Hemisphere mean temperature series provided by Mann et al. (2009) as a reasonable choice of an uncertainty estimate. Accordingly, we inflate the ensemble by a random realisation drawn from a distribution with standard deviation equal

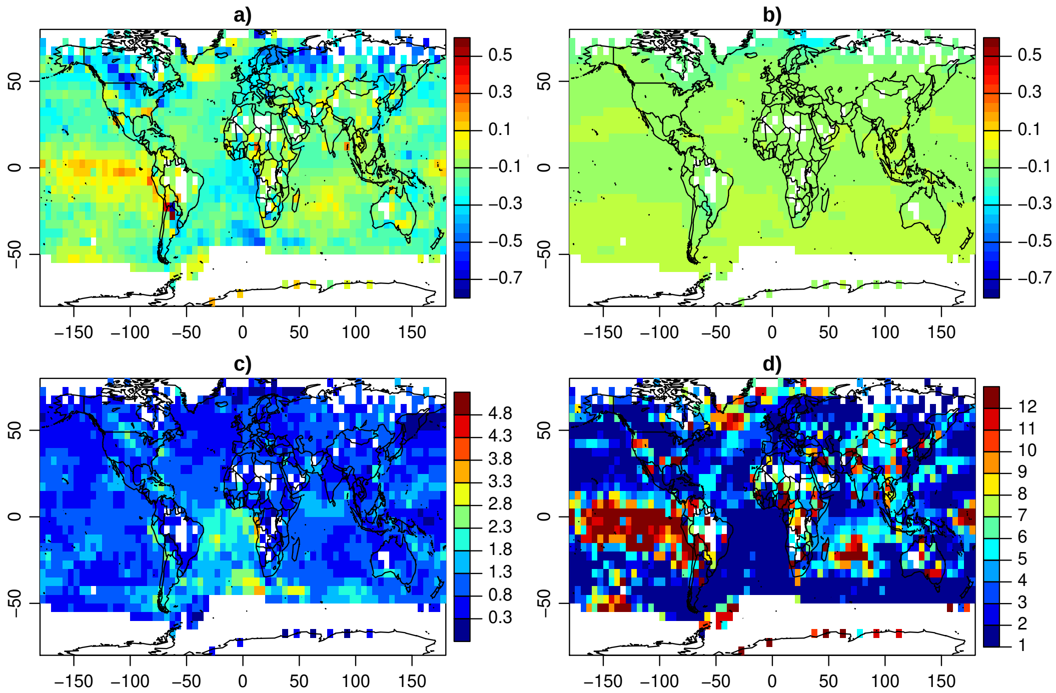

Fig. 7. Global fields of decadally smoothed temperature: **(a)** reconstructed mean anomaly map for a cold period (for the 1390s to 1690s), **(b)** ensemble-mean simulated anomaly map for the same period, **(c)** ensemble mean of relative standard deviations (reconstruction standard deviation divided by simulation standard deviation at each grid-point for the full period), **(d)** mapped target ranks for evaluating SIM against the target for the cold period (1390s to 1690s).

to this standard error. Without uncertainty inflation, the expected effective rank frequencies can be very small due to the temporal auto-correlations 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).

3.2.1 Ensemble consistency of spatial fields

Figures 8 to 10 display a selection of results for the evaluation of consistency of the SIM ensemble with the global temperature field reconstruction by Mann et al. (2009) at individual grid points. As for the time series data, the most common deviation is a too wide ensemble. This holds for the probabilistic assessment via rank counts in Fig. 8 (for a random selection of grid points) and for the climatological evaluation via residual quantiles in Fig. 9 (for the same selection of grid points). However, we also find grid points where the rank counts suggest an under-dispersive, too narrow ensemble. These are mostly due to opposite probabilistic biases. There are also grid points at which flat rank counts do not allow to reject consistency over sub-periods and over the full period. Again, full-period consistency may be due to opposite biases in different sub-periods.

There are notable shifts in the rank counts between sub-periods (Fig. 8). That is, consistency changes over time. Opposing biases are especially prominent, and the SIM ensemble is moderately (or even extremely) biased in at least one sub-period.

The climatological residuals highlight even more strongly the lack of consistency between the ensemble and the target (Fig. 9). Deviations from the target are similar for the individual SIM ensemble members. The prominent slopes in residual quantiles highlight the stronger variability in SIM even for decadal moving averages. At certain grid points, however, the analyses suggest under-dispersion or even consistent climatologies. Differences in residuals between sub-periods are diverse but can be rather small between the first and the last 250 records (compare Fig. 9). Residuals can be small or even nearly vanish in the last sub-period. However, there are also grid points where biases increase, change sign or where deviating spread characteristics become more pronounced. Furthermore, target and SIM distributions, or both, may be completely different between the first and the last sub-period (compare Fig. 9). This complies with the shifts in the probabilistic analyses (Fig. 8). Thus results are often not comparable between sub-periods either probabilistically or climatologically. The subsequent shifting emphasises the general lack of a common signal, and specifically, differences in the long-term trend component.

The decadal smoothing of the target data reduces the width of the climatological quantile distributions, and a number

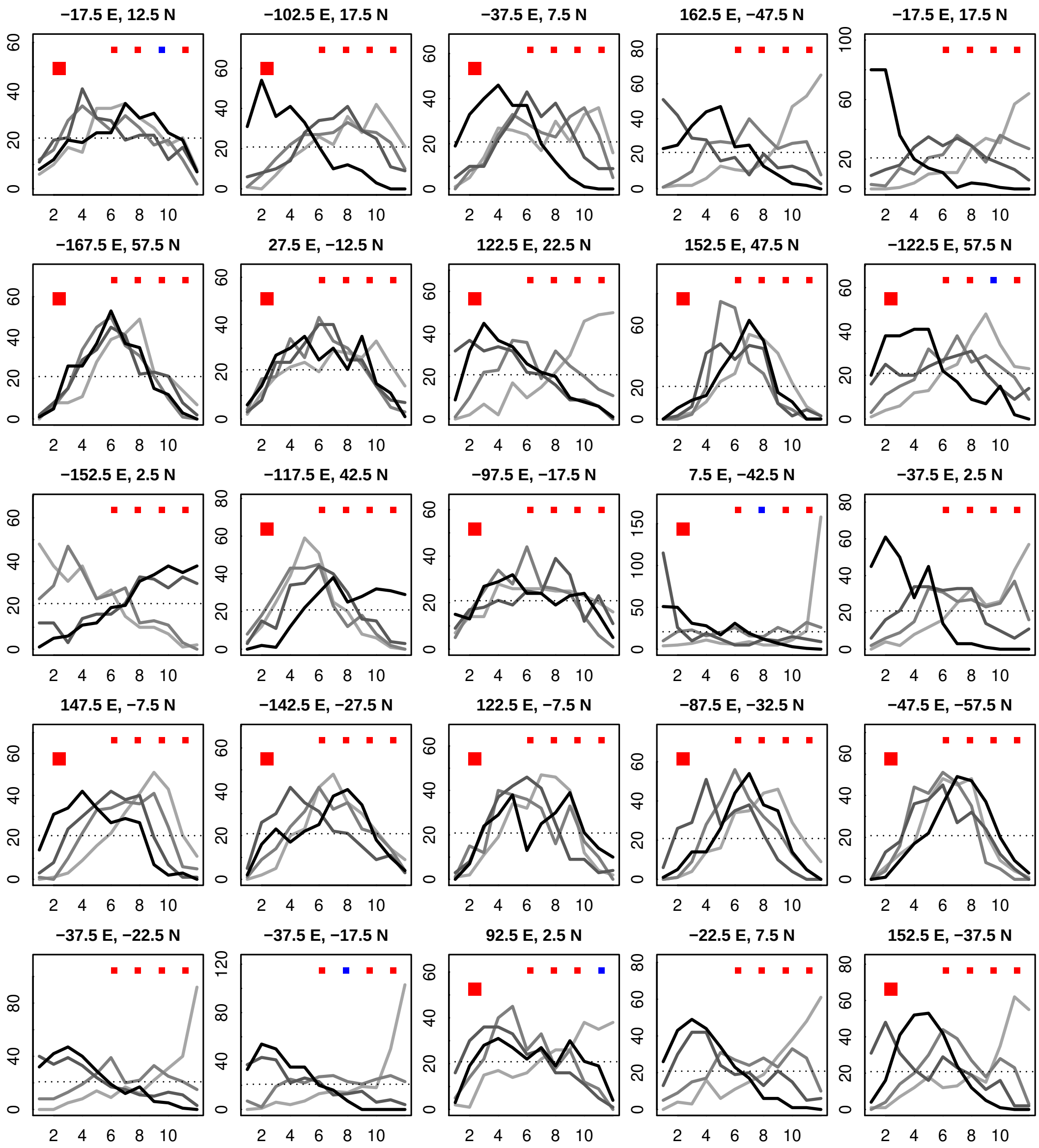

Fig. 8. Rank histogram counts for a random selection of 25 grid points from the decadal smooth global temperature data and the first, second, third and last 250 non-overlapping records of the decadally smoothed annual data (grey to black lines, about 800 to 1050, 1050 to 1300, 1300 to 1550, and 1595 to 1845). Large (small) red squares mark grid points where spread or bias deviations are significant over the full period (the individual sub-period). Blue squares 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.

of grid points display only very small quantile-ranges due to very weak inter-decadal variability (not shown). Narrow quantile distributions of the target result in particularly strong climatological over-dispersion at certain grid points. The target and ensemble quantile distributions can be broader in higher Northern Hemisphere latitudes than at other locations.

The selection of grid points in Figs. 8 and 9 provides only a snapshot of the results for the global temperature field data. Figure 10 summarises the full and single-deviation goodness-of-fit test statistics for the full period and the sub-periods defined above. We account for the target uncertainties in all displayed results. We use a moderate random estimate to account for the target uncertainties ($\sigma = 0.1729$, see Sect. 2.2).

In Fig. 10, the red colour marks areas where the tests indicate that deviations from a uniform rank count are significant at the 90 % level. The SIM data is probabilistically consistent with the target over spatially extended regions only over Central Eurasia and the tropical Pacific for the full period according to the full test (Fig. 10a). Results diverge over

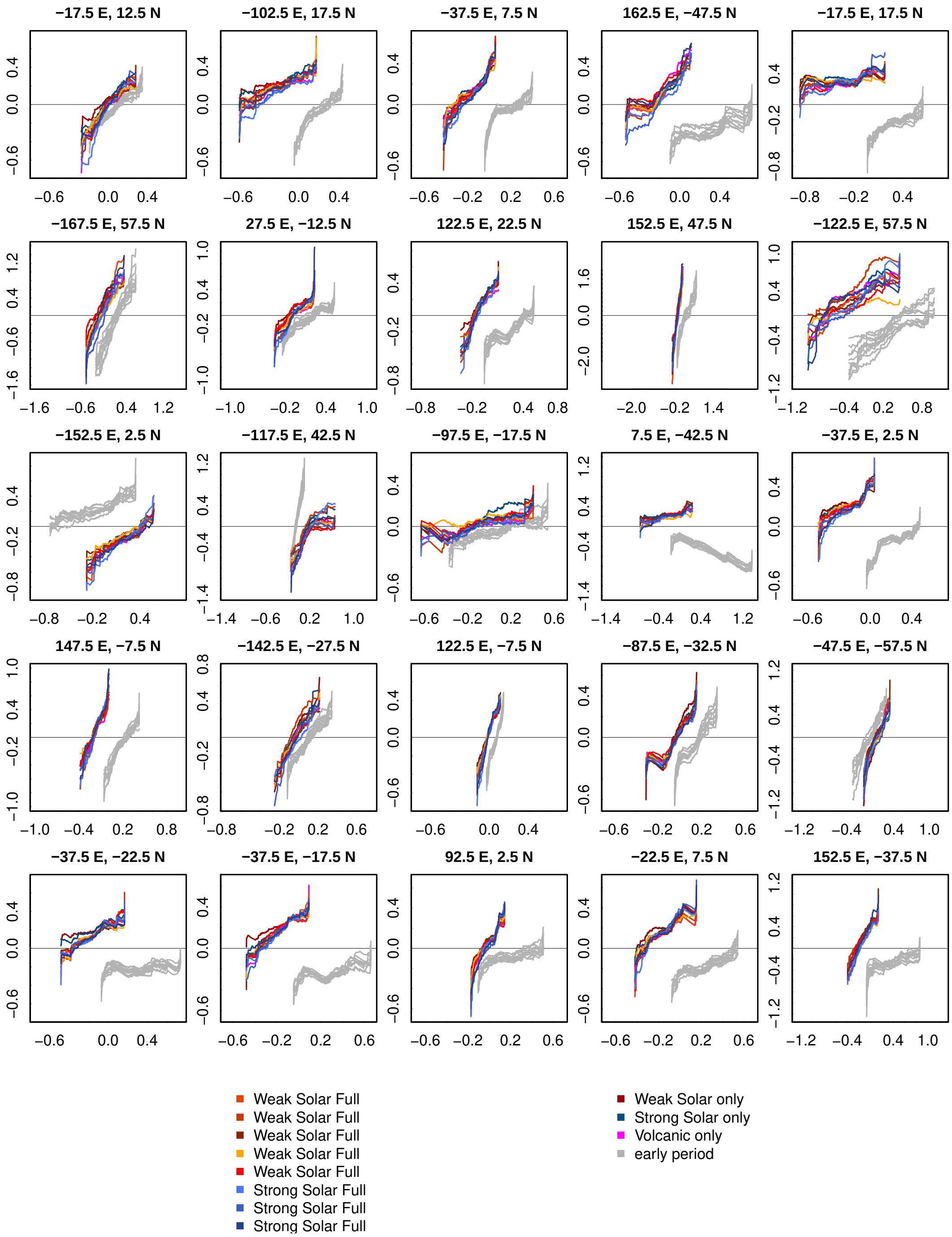

Fig. 9. Residual quantile-quantile plots for a random selection of 25 grid points from the decadal smooth global temperature data and the first (grey) and the last (colours) 250 records. Locations given in titles of individual panels. For representation see legend.

the sub-periods of 250 records. For example, SIM is consistent with the target in the North Atlantic sub-polar gyre region for the early sub-period (about 800 to 1050, Fig. 10b) but not for the following one (Fig. 10c). Overall, opposite results are common in the full test for these early two periods. SIM is consistent with the target over wide regions of Eurasia and North America in the latter but not in the early one. Subsequently, the ensemble appears to be consistent with the target in northern North America, the tropical Pacific and south of Greenland during the sub-period from about 1300 to 1550 (the early Little Ice Age, Fig. 10d). In the last period (Fig. 10e, about 1595 to 1845), the full

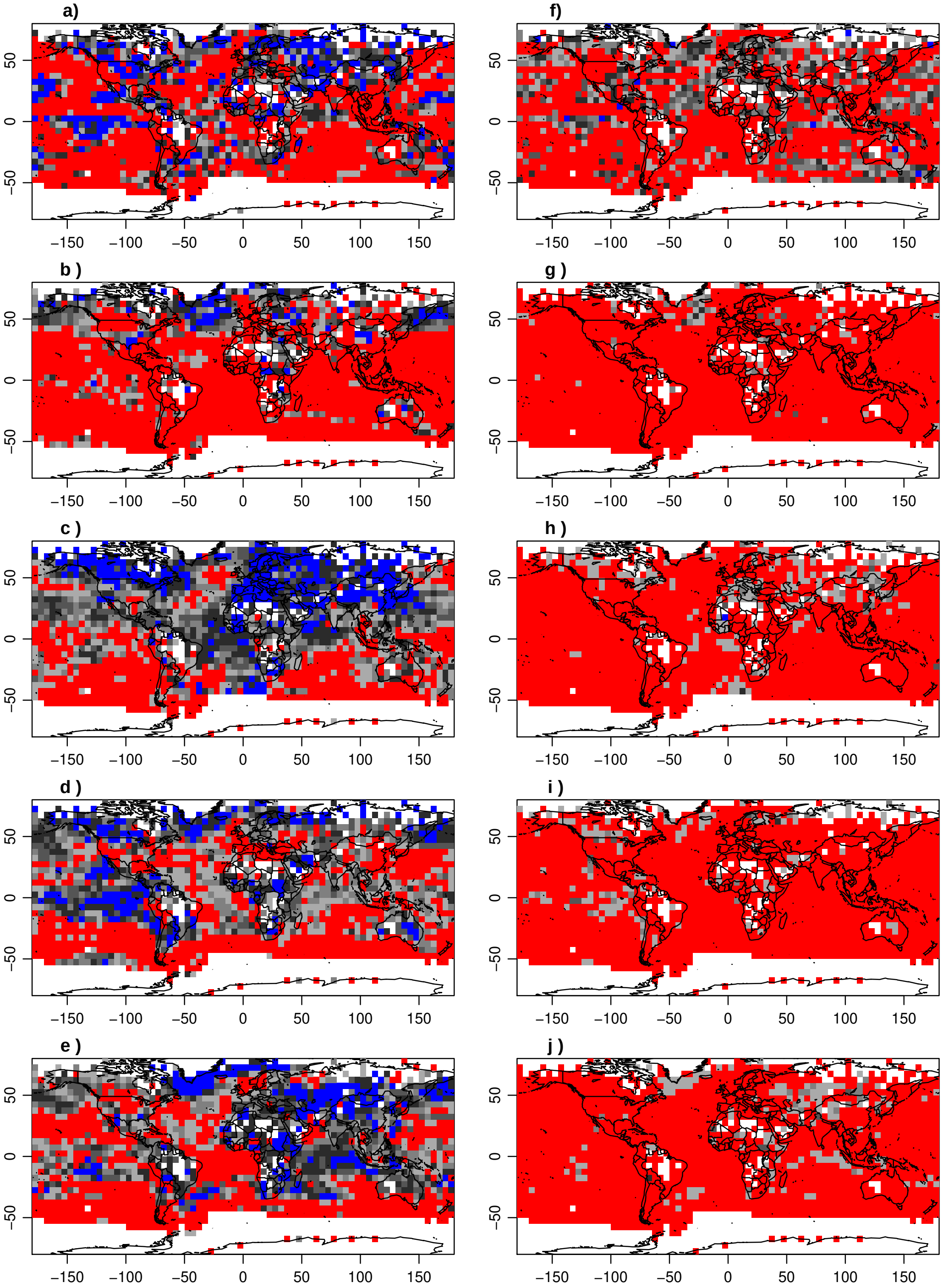

Fig. 10. Global assessment of the goodness-of-fit test for the decadal smooth data considering uncertainties in the target. Plotted are lower p values. In the left column: full χ^2 test, in the right column: maximum of p values for single deviation tests for bias and spread. Blue for smaller than 0.1, dark to light grey in steps of 0.2 for the range between 0.1 and 0.9, red for 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.

test again suggests that the SIM ensemble is consistent with the target over Eurasia and the North Atlantic according to the full test. On the other hand, single deviations are nearly always and everywhere significant (Fig. 10f–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 we consider the accumulated field data over all data points in space and time and if we account for the target-uncertainties, we again find notable probabilistic over-dispersion (not shown) and also the climatological assessment indicates general over-dispersion (not shown).

In summary, even more prominent than for the area-averaged time series, the SIM ensemble displays a lack of consistency with its target for decadally smoothed global temperature field data. The different diagnosed biased, under- and over-dispersive discrepancies suggest that the relations between ensemble and target differ strongly in different regions. However, we cannot reject a uniform outcome of the rank counts for some regions and certain periods based on the full test. This may be to some extent due to a very small number of independent samples. Lack of consistency is most prominent over the southern oceans. Tests for the single deviations of bias and spread are nearly everywhere significant. Thus general consistency between the SIM ensemble and its field data target remains very weak. Note that the (lack of) consistency is not homogeneous in time, but may differ between selected periods. It is not necessarily valid to assume an increase in consistency with decreasing temporal distance to the present.

3.2.2 Addressing lack of consistency in spatial fields

Figure 7a presented the reconstructed mean anomaly map for an arbitrary sub-period (1390s to 1690s) encompassing part of the Little Ice Age (LIA). The LIA was chosen as it depicts one period of special interest in the literature. This period is only partially captured in the previous assessments of consistency. However, Fig. 10 indicates that we cannot expect too many differences between the considered sub-periods and this partially independent one. Based on Fig. 7, we are going to trace possible sources of lack of consistency.

The reconstructed estimate basically fully relies on a statistical relation between observations and the proxies. The simulated estimate relies on our knowledge on the physics of the climate system as coded in the simulator.

We note that the amplitudes of mean anomalies are comparable between reconstructions and SSIM simulations except in the tropical Pacific, but the WSIM ensemble members display less cooling in the selected period (not shown, compare reconstruction in Fig. 7a, SIM ensemble mean in Fig. 7b and rank map in Fig. 7d). Mapped ranks in Fig. 7d exemplify the potential differences in simulated and reconstructed mean anomaly patterns. Obviously, there are large discrepancies between both approaches as highlighted by the cold bias of the SIM ensemble in the tropical Pacific and further spatially extended biases in most oceanic regions. SIM is biased low over the tropical Pacific ocean but a high bias is seen over most other oceanic regions, North America and eastern and western Eurasia. These biases are not representative for the full period as we discuss above (compare Fig. 10). Rather, Fig. 7c highlights how strongly mean anomalies of SIM may disagree with the target patterns for specific periods.

On the other hand, variability is often regionally comparable over the full period of the data (compare the ensemble-mean relative standard deviations in Fig. 7c) but also over sub-periods (not shown). Nevertheless, the reconstructed variability is strongly overestimated in the South Atlantic or more generally southern hemispheric ocean regions. Similarly, the SIM members often display more variability than the target over the other oceanic regions (see Fig. 7c). Sub-periods give comparable patterns, but slight changes may of course be found in the specific size of over or underestimation of variations. Differences in variability between target and ensemble are rather small-scale over the continents. The rank counts in Fig. 8 reflect these regional differences in variability. For instance, grid points in the southern hemispheric Atlantic sector suggest opposite biases in different periods, while, for the mid- to high-latitude North Pacific grid points, they suggest largest over-dispersion for the model-ensemble.

The mapped ranks (Fig. 7d) highlight another feature that also appears in other sub-periods and even for some further field reconstructions (not shown): the reconstruction target generally represents the largest absolute mean anomalies in the set of SIM ensemble and target.

Thus reconstructions and simulations commonly differ in the mean and in the variability for certain periods. The SIM ensemble generally underestimates the size of the mean anomalies with reconstructed warm anomalies being warmest and cold anomalies coldest, which results in ensemble biases. Further, the ensemble members vary more strongly in the averaging periods, which leads to common over-dispersive relations. The latter feature is amplified in the analyses of consistency by considering the uncertainty of the target. The underestimating biases possibly relate to general differences in the long-term trend between the ensemble members and the target field reconstruction. These, in turn, are spatially explicit expressions of the differences in the long-term tendencies that were similarly found in the large-scale mean data (compare Figs. 2 and 3). Franke et al. (2013) report a general overestimation of low-frequency variability in proxies and reconstructions. This, in turn, possibly explains our finding of more variability in the simulations, also on the decadal scale, compared to the reconstruction.

Both, differences in trend and in variability, express a general misrepresentation of the climate statistics. Therefore, comparing anomaly patterns is of reduced value due to a general dissimilarity between reconstructions and simulations. Thus the assessment of ensemble consistency not only reduces the subjectivity of a comparison between simulations and reconstructions but, in turn, may help in clarifying sources of disagreement in the statistics.

4 Discussions of the results

We realign the simulations and the reconstructions to the mean of a common period to correct systematic differences in long-term trends before applying tests of consistency (similar to traditional simulation-reconstruction comparisons, e.g. Jansen et al., 2007; Brázdil et al., 2010; Luterbacher et al., 2010; Jungclaus et al., 2010; Zorita et al., 2010; Zanchettin et al., 2013). For instance, 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. They also highlight periods in which the simulations are rather warm compared to temperature reconstructions when temperature deviations are considered with respect to the period 1961–1990 (e.g. in the 12th and 13th centuries). We accept that the choice of the reference period influences the results.

Strong probabilistic and climatological deviations arise, in some cases, between the considered ensembles of simulations and reconstructions for the hemispheric data. Results are to some extent dependent on the utilized uncertainty estimates and the reference periods. For the Northern Hemisphere data, the choice of the specific sub-ensemble also has an influence. The simulation ensemble is also generally over-dispersive for the seasonal European temperature reconstructions by Luterbacher et al. (2002, 2004) and Xoplaki et al. (2005) or the South American austral summer temperature reconstructions by Neukom et al. (2011) as targets (not shown). Even if the ensemble is consistent according to our analyses at the grid-point level or for area-averaged index-series, the associated uncertainties usually lessen the value of such consistency. Only the annual Central European temperature time series data arises as fully consistent between the simulation ensemble and the reconstruction. Thus the SIM ensemble is only consistent with an estimate for the last 500 yr and, therefore, may benefit from a more stable number of reliable available proxy indicators compared to longer period reconstructions. The forcing data used to drive the simulations can also be assumed to be less uncertain in this shorter period compared to the full millennium. However, 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). On the other hand, Franke et al. (2013) highlight the general overestimation of low-frequency variability in proxies and reconstructions compared to observations and simulations.

As noted in Sect. 2.3, it is convenient, but not necessarily appropriate, to employ the raw ensemble reconstructions by Frank et al. (2010) as representing inter-annual variations. Similarly, it is arguable whether or not an ensemble mean represents inter-annual variability. Results change notably when uncertainties are included or excluded and/or when internal variability in the assessment of the FRS ensemble against the target of the SIM ensemble mean is considered. Although the temporal evolutions notably deviate, it appears likely that the FRS and most of its members are indeed consistent with the target of the SIM ensemble mean under the assumptions made on internal variability and the uncertainties. On the other hand, the SIM ensemble displays pronounced deviations from consistency relative to the target of the FRA ensemble mean including different 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 (compare Figs. 1–3). The 20th century disagreement is possibly due to the evolution of the simulations within the SSIM ensemble (i.e. with strong solar forcing). The considerations on internal variability introduce an additional source of uncertainty. While the consideration of internal variability reduces the problems in employing ensemble-mean targets, it also highlights the ambiguity of our estimates of past climate trajectories.

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 spatiotemporal distribution of proxy series. Their goal is similar to the approach utilized here. Their framework focuses on the similarity between simulated and reconstructed series by analysing two newly developed correlation-based and distance-based test statistics. Hind et al. (2012) apply their approach in a pseudo-proxy experiment within the virtual reality of the COSMOS-Mill sub-ensembles to 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 under the paradigm of statistical indistinguishability.

Finally, 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. This also allows for a detailed assessment of the consistency between simulations and reconstructions. Preliminary analyses for the available CMIP5/PMIP3-past1000 simulations indicate that, for the European and northern hemispheric temperature reconstructions considered in the present study, the multi-model ensemble behaves similar to the COSMOS-Mill ensemble with respect to probabilistic and climatological consistency.

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 a verification data set (e.g. Annan and Hargreaves, 2010; Marzban et al., 2011). If no reliable observable target can be identified, as is the case in periods and regions without instrumental observations, such statistical analyses reduce the subjectivity in comparing 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 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.

We considered the COSMOS-Mill ensemble (Jungclaus et al., 2010) and various reconstructions within the described approach. We found the simulation ensemble to be consistent, within sampling variability, with the Central European temperature reconstruction by Dobrovolný et al. (2010). 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 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 possibly consistent relative to the simulation ensemble mean.

Furthermore, the simulation ensemble is found to be statistically distinguishable from the global field temperature reconstruction by Mann et al. (2009). Although the data is probabilistically consistent for multi-centennial sub-periods and certain regions according to the applied full test, analyses of single probabilistic deviations and climatological differences emphasise a general lack of consistency. We found the largest, but still limited, 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 ensemble and reconstructions stresses the importance of improving simulations and reconstructions to investigate past climates in order to achieve a more resilient estimate of the 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 this simulation ensemble 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, which have been described in the present work.

Appendix A

Evaluation of the rank histograms

The goodness-of-fit χ^2 statistics and the respective p values depend on the degrees of freedom of the distribution (see Jolliffe and Primo, 2008). The distributional degrees of freedom equal $n-1$ for the full test and n is the number of classes in the rank histogram. The decomposition of the χ^2 test statistic implies that we have only 1 degree of freedom for the single deviation test (Jolliffe and Primo, 2008; Annan and Hargreaves, 2010).

We reject consistency for certain right p values of the test. Where appropriate, we also interpret the test statistics in terms of a reversed null hypothesis to test 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 chi-square values become 2.706 for the single deviation test. For the full goodness-of-fit test, we consider ensembles of eleven, nine, five and three members (see Sect. 2.2). Critical values are respectively 17.275, 14.684, 9.236 and 6.251.

Meaningful results for the tests require accounting for dependencies in the data (Jolliffe and Primo, 2008; Annan and Hargreaves, 2010). All analyses account for effective sample size (see discussions by and references of Bretherton et al., 1999). A larger effective sample size essentially leads to a higher chance of rejecting the hypothesis of uniformity. Furthermore, the results are sensitive to the made assumptions, particularly those with respect to the included uncertainty estimates (see Sect. 2.3).

Some further notes are in place. 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). Temporal correlations further affect the structure of the rank histograms (Marzban et al., 2011; Wilks, 2011), and sampling variability can result in erroneous conclusions from the rank counts. That is, a flat rank histogram is only a necessary condition for consistency (see discussions by e.g. Hamill, 2001; Marzban et al., 2011). To account for this, we display, for area-averaged time series, quantile statistics of block-bootstrapped rank histograms (Marzban et al., 2011; 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. This additionally 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.

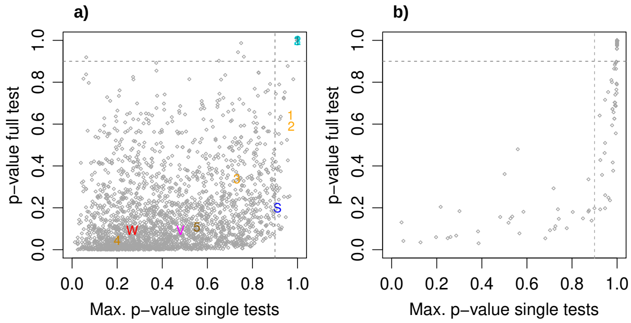

Fig. B1. Surrogate ensemble (SUR): **(a)** Testing against 2201 surrogate targets: p values for the full goodness-of-fit χ^2 test plotted against the maximum of the p values obtained from the tests for bias and spread deviations. In **(a)** orange (blue, see top right corner) numbers 1–5 (1–3) give values for the WSIM (SSIM) 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 in **(a)** but for the test against the 521 members of the Frank et al. (2010) ensemble (FRA) as targets. Horizontal and vertical lines indicate a conservative 90 % level for significance against the null hypothesis of a uniform rank histogram. **(b)** accounts for uncertainties and reduced internal variability in data by Hegerl et al. (2007), Mann and Jones (2003) and Mann et al. (2008).

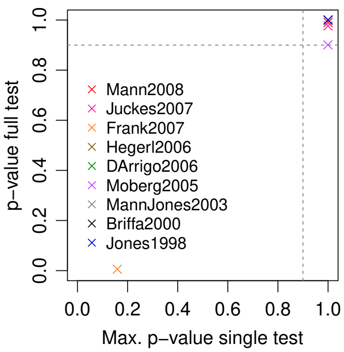

Fig. B2. The Northern Hemisphere reconstruction sub-ensemble re-calibrated to 1920–1960 (FRS): test for consistency of the remaining members of the ensemble against a target defined by one of the members: p values for the full goodness-of-fit χ^2 test plotted against the maximum of the p values obtained 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.

Appendix B

Intra-ensemble consistency

We shortly describe the within-ensemble consistencies. Therefore, we construct a surrogate simulation ensemble (SUR) of eleven 850 yr long series from the 3100 yr of the control run. We further use 2201 segments of the control run as potential verification targets. The number is arbitrarily chosen. SUR is probabilistically consistent with these surrogate targets as well as with three of the weak solar full-forcing simulations, the weak solar forcing only simulation and the volcanic forcing only simulation. The full goodness-of-fit tests allow rejecting uniformity in less than one percent of the surrogate targets according to the test statistics (see Fig. B1a) thereby indicating general consistency of SUR with the surrogate targets. The single deviation tests are significant in less than 50 cases (see Fig. B1a). Here, we do not include uncertainty estimates. Thus an ensemble of unperturbed simulated climate estimates is consistent with at least some simulated forced climates.

Section 3.1 considers the ensemble mean of the FRS ensemble of Northern Hemisphere temperature reconstructions. Since the reconstructions notably differ from one another (compare Frank et al., 2010), we may question the consistency of the ensemble with each member. Here, we consider the target uncertainty and account for the reduced internal variability in the filtered time series by Hegerl et al. (2007), Mann and Jones (2003) and Mann et al. (2008). The FRS ensemble is only probabilistically consistent with respect to the recalibrated Frank et al. (2007) reconstruction (Fig. B2).

Next, we consider all 521 members of the FRA ensemble as potential targets for the surrogates. We include uncertainty estimates and compensate for reduced internal variability in the filtered reconstructions by Hegerl et al. (2007), Mann and Jones (2003) and Mann et al. (2008). The SUR ensemble is consistent with about 14 % of the FRA members according to the full test although they do not include a common signal (Fig. B1b). Test statistics for the single tests are not significant in about 7 % of the cases (Fig. B1b). Similarly, the climatological analyses indicate larger consistency for the surrogate ensemble than for the real ensemble (not shown). That is, the simulated forced climate evolutions may differ very strongly from some of the reconstructed targets (not shown), and the unperturbed internal climate variability may be indistinguishable from forced simulated or reconstructed variability.

Thus we generally cannot reject consistency for an ensemble and its verification data (Fig. B1a) if the 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). Similar considerations in seasonal and medium-range weather forecasting (Johnson and Bowler, 2009) depict that ensembles are consistent as long as the target variability and the projected variability are similar. The FRS reconstruction ensemble apparently does not generally comply with these assumptions (Fig. B2).

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. Three 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 the University of Hamburg.

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

Edited by: V. Rath

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, doi: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, doi: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., Laîné, A., Loutre, M.-F., Marti, O., Merkel, U., Ramstein, G., Valdes, P., Weber, S. L., Yu, Y., and Zhao, Y.: Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum – Part 1: experiments and large-scale features, Clim. Past, 3, 261–277, doi:10.5194/cp-3-261-2007, 2007.

Braconnot, P., Harrison, S. P., Kageyama, M., Bartlein, P. J., Masson-Delmotte, V., Abe-Ouchi, A., Otto-Bliesner, B., and Zhao, Y.: Evaluation of climate models using palaeoclimatic data, Nat. Clim. Change, 2, 417–424, doi:10.1038/nclimate1456, 2012.

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

Bradley, R. S.: High-resolution paleoclimatology, in: Dendroclimatology, edited by: Hughes, M. K., Swetnam, T. W., and Diaz, H. F., Developments in Paleoenvironmental Research, Vol. 11, chapter 1, Springer, Dordrecht, 3–15, doi: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, doi: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, Quaternary Sci. Rev., 19, 87–105, doi:10.1016/S0277-3791(99)00056-6, 2000.

Crowley, T. J. and Unterman, M. B.: Technical details concerning development of a 1200-yr proxy index for global volcanism, Earth Syst. Sci. Data Discuss., 5, 1–28, doi: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, doi: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, doi: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, doi:10.1029/2006PA001304, 2007.

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

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

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

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

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

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, doi: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, doi: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 – Part 2: A pseudo-proxy study addressing the amplitude of solar forcing, Clim. Past, 8, 1355–1365, doi:10.5194/cp-8-1355-2012, 2012.

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

Mann, M. E. and Jones, P. D.: Global surface temperatures over the past two millennia, Geophys. Res. Lett., 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, P. Natl. Acad. Sci., 105, 13252–13257, doi: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, doi:10.1126/science.1177303, 2009.

Marzban, C., Wang, R., Kong, F., and Leyton, S.: On the effect of correlations on rank histograms: reliability of temperature and wind speed forecasts from finescale ensemble reforecasts, Mon. Weather Rev., 139, 295–310, doi:10.1175/2010MWR3129.1, 2011.

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

Misios, S. and Schmidt, H.: Mechanisms involved in the amplification of the 11-yr solar cycle signal in the Tropical Pacific Ocean, J. Climate, 25, 5102–5118, doi: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, doi: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, doi: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.

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

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

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

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, 8 pp., doi:10.1051/0004-6361/201016173, 2011.

Steinhilber, F., Beer, J., and Fröhlich, C.: Total solar irradiance during the Holocene, Geophys. Res. Lett., 36, L19704, doi: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, Clim. Past, 8, 1339–1353, doi:10.5194/cp-8-1339-2012, 2012.

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

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, doi:10.1175/2010MWR3446.1, 2011.

Wilson, R., D'Arrigo, R., Buckley, B., Büntgen, U., Esper, J., Frank, D., Luckman, B., Payette, S., Vose, R., and Youngblut, D.: A matter of divergence: tracking recent warming at hemispheric scales using tree ring data, J. Geophys. Res., 112, D17103, doi:10.1029/2006JD008318, 2007.

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

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

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., 40, 1301–1318, doi:10.1007/s00382-012-1361-9, 2013.

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