Clim. Past, 8, 1355–1365, 2012
www.clim-past.net/8/1355/2012/
doi:10.5194/cp-8-1355-2012

Climate of the Past

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

A. Hind[1], **A. Moberg**[1], **and R. Sundberg**[2]

[1]Department of Physical Geography and Quaternary Geology, Bert Bolin Centre for Climate Research, Stockholm University, 106 91 Stockholm, Sweden
[2]Department of Mathematics, Division of Mathematical Statistics, Stockholm University, 106 91 Stockholm, Sweden

Correspondence to: A. Hind (alistair.hind@natgeo.su.se)

Received: 13 December 2011 – Published in Clim. Past Discuss.: 12 January 2012
Revised: 11 June 2012 – Accepted: 6 July 2012 – Published: 27 August 2012

Abstract. The statistical framework of Part 1 (Sundberg et al., 2012), for comparing ensemble simulation surface temperature output with temperature proxy and instrumental records, is implemented in a pseudo-proxy experiment. A set of previously published millennial forced simulations (Max Planck Institute – COSMOS), including both "low" and "high" solar radiative forcing histories together with other important forcings, was used to define "true" target temperatures as well as pseudo-proxy and pseudo-instrumental series. In a global land-only experiment, using annual mean temperatures at a 30-yr time resolution with realistic proxy noise levels, it was found that the low and high solar full-forcing simulations could be distinguished. In an additional experiment, where pseudo-proxies were created to reflect a current set of proxy locations and noise levels, the low and high solar forcing simulations could only be distinguished when the latter served as targets. To improve detectability of the low solar simulations, increasing the signal-to-noise ratio in local temperature proxies was more efficient than increasing the spatial coverage of the proxy network. The experiences gained here will be of guidance when these methods are applied to real proxy and instrumental data, for example when the aim is to distinguish which of the alternative solar forcing histories is most compatible with the observed/reconstructed climate.

1 Introduction

Variations of solar irradiance on long time scales have a potential influence on global climate. Instrumental satellite-based measurements of total solar irradiance (TSI) are, however, available only back to the mid-1970s. Within this period, TSI monitors show an 11-yr cycle with an amplitude of about 0.07 %, in phase with the sunspot number cycle. To estimate TSI further back in time, several investigators have relied on observed correlations between various indices of solar activity in combination with assumptions of how these indices are related to variations in TSI (see Gray et al., 2010, for a thorough review).

One of the most highly debated questions concerns whether there exists a centennial-scale variation in the background level of TSI. Different estimates of the background amplitude of TSI are often characterized by their hypothesized decrease in TSI values within the Maunder Minimum (MM) period of low solar activity, during 1645–1715 AD, compared to the recent satellite-based measurements. Estimates made in the 1990s suggested rather large values between 0.24 % and as much as 1 % (Reid., 1991; Hoyt and Schatten, 1993; Lean et al., 1995; Zhang et al., 1994; Reid., 1997; Cliver et al., 1998; Bard et al., 2000). Continued research in the 2000s (Wang et al., 2005; Krivova et al., 2007; Tapping et al., 2009; Steinhilber et al., 2009; Gray et al., 2010) did not support these results; the most widely accepted view now is that the background variations are between 0.04 % and 0.1 %, which is the range adopted by the Paleoclimate Model Intercomparison Project Phase III (PMIP3)

(Schmidt et al., 2011). To put these different estimates into context, a change in TSI by 0.1 % corresponds to a radiative forcing that is about one-tenth of the current anthropogenic forcing from greenhouse gases (Lockwood, 2011). The debate, however, is not yet over. Very recently, two author teams challenged the currently held view, where one team (Shapiro et al., 2011) hypothesized that the decrease at MM could be more than 0.4 %, while the other team (Schrijver et al., 2011) argued that there could possibly be no change at all.

One way to attempt constraining the long-term amplitude of solar forcing is to use alternative TSI histories to drive climate model simulations, and then see which forcing history provides simulated temperatures that are most compatible with the observed past temperatures and reconstructed past temperatures derived from proxy data (Ammann et al., 2007; Jungclaus et al., 2010; Feulner, 2011; Schmidt et al., 2011). This approach, however, is associated with difficulties because of the always present noise in the climate proxy data (Jones et al., 2009) in combination with the stochasticity of the internal (unforced) variability of the climate system (Yoshimori et al., 2005). Another complicating factor is uncertainty regarding the Earth's climate sensitivity to radiation changes and the varying climate sensitivity among different climate models (Knutti and Hegerl, 2008). These difficulties provide a motivation for the experiment we undertake here, which is designed such that we define "true" temperatures derived from simulations with a single climate model, where we know with certainty what the amplitude of solar forcing has been and that the climate sensitivity issue can be ignored. Moreover, we know precisely how much noise there is in our proxy data, because they are constructed from simulated "true" temperatures but with known noise added. We then ask the following: Given knowledge of the true solar forcing, the true past temperatures, and the level of proxy noise, is it possible to determine whether a forced simulation with a climate model, which includes the correct solar forcing amplitude, gives a smaller distance to the reconstructed temperatures than expected from a control simulation with constant forcings? And, if so, can we correctly rank simulations driven by the correct TSI amplitude, such that they are deemed better than other simulations that include an alternative incorrect amplitude?

A study of this kind is a variant of a now common approach in paleoclimatology, known as a pseudo-proxy experiment, where output from climate model simulations is used to test the performance of different methods to reconstruct past climates (see Smerdon, 2012, for a review). In our pseudo-proxy study, we use the newly developed statistical framework of our companion paper (Sundberg et al., 2012; henceforth referred to as Part 1) to rank or distinguish between model simulations using two different solar forcings, either as single forcings or in conjunction with other important forcings used in tandem. Note that we do not attempt to address the question of whether a higher or lower solar variability imposed on simulations is closer to reality. We merely state that the issue is of great importance and choose it as a focal subject in the testing of our framework's sensitivity. Ultimately, this will allow better judgement regarding how possible it is, in future comparisons, to identify which simulation is best able to simulate observed temperatures in real proxy and instrumental data. As our pseudo-proxy experiment test-bed, we use the set of simulations from the Community Earth System Modeling (COSMOS) Millennium Activity of the Max Planck Institute (Jungclaus et al., 2010).

2 The COSMOS Millennium Activity – model description and experimental design

The COSMOS Millennium Activity simulation experiments were conducted using the Max Planck Institute Earth System Model (MPI-ESM), which is formed from an atmospheric model ECHAM5 (Roeckner et al., 2003), an ocean model MPIOM (Marsland et al., 2003) and models for both land vegetation (JSBACH) and ocean biogeochemistry (HAMOCC). The model resolution is T31 (3.75°) for ECHAM5, and MPIOM applies a conformal grid with a horizontal resolution ranging from 22 km to 350 km (Jungclaus et al., 2010). The ocean and atmosphere are coupled daily without flux correction.

The Millennium Activity involved the creation of a 3000-yr unforced control (CTRL) simulation, after a multi-century spin-up phase in which the carbon cycle was brought into equilibrium. The CTRL model experienced 800 AD orbital conditions and pre-industrial greenhouse gas concentrations (Jungclaus et al., 2010). In our experiment, it was separated into three 1000-yr-long CTRL simulations to be used in the comparison with the forced simulations. The globally averaged land-only annual temperature anomalies (30-yr means) of the three CTRL simulations are shown in Fig. 1a. To account for some of the previously discussed uncertainty in the magnitude of solar forcing, the Millennium Activity conducted experiments using both "low" and "high" estimated TSI forcing series. The "low" forcing exhibits a total TSI reduction of 0.1 % at the Maunder Minimum compared to the present (Krivova et al., 2007 reconstruction – in agreement with the largest amplitude used in PMIP3) against a forcing with a "high" reduction of 0.25 % (Bard et al., 2000 reconstruction, representative of a common late-1990s view). Other forcings known to be principal drivers of climate were also included in the experiments: orbital, volcanic and non-volcanic aerosols, greenhouse gases (CO_2, CH_4, N_2O), as well as land-use changes (see Jungclaus et al., 2010, for details).

Two full-forcing ensembles, representing the last 12 centuries, were generated by starting simulations from different ocean initial conditions and are separated by their respective "low" E1 (Fig. 1b, five simulations) and "high" E2 (Fig. 1d, three simulations) solar forcing histories, as well

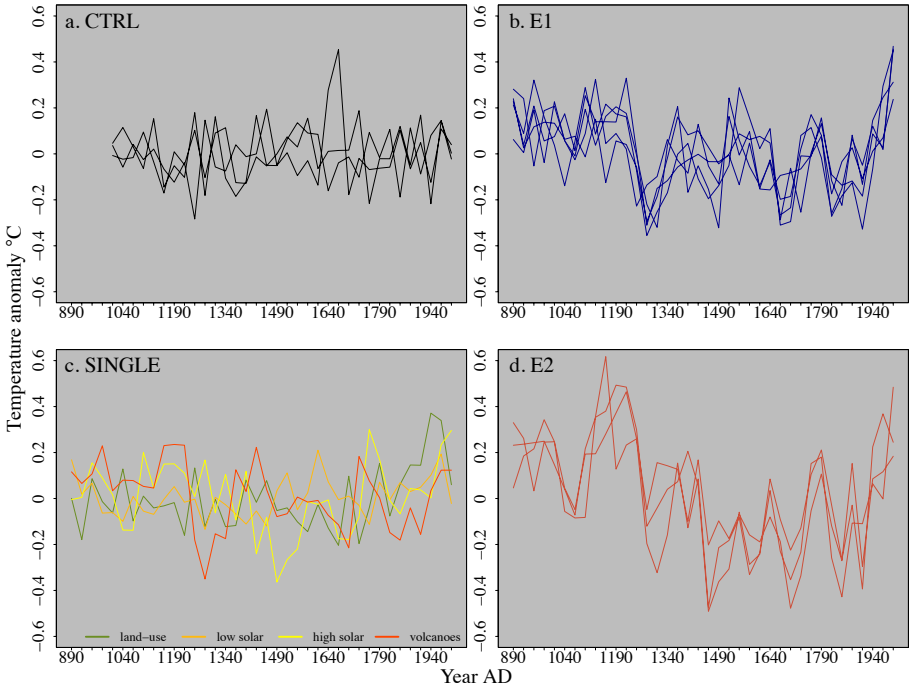

Fig. 1. The MPI Millennium Activity COSMOS simulations over the last millennium with 30-yr non-overlapping means of global land-only annual temperature anomalies (°C) from the period 850–2000. The simulations are shown as the CTRLs (top-left panel), E1 ensemble (top-right panel), SINGLE forcing (bottom-left panel) and E2 ensemble (bottom-right panel). The SINGLE forcing simulation series are land-use changes (green), low solar (light orange), high solar (yellow) and volcanoes (red).

as any solar-induced CO_2 concentration changes (which are possible through the model's interactive carbon cycle). A representation of the forcings is shown in Fig. 2. Note that these single time series representations of the global forcings are shown in terms of their annual mean radiative forcing at the top of the atmosphere. In addition to the two full-forcing simulation ensembles, the model was also driven by each forcing individually to create several single-forcing simulations (Fig. 1c). There is a pronounced simulated warming in the 20th century associated with the enhanced greenhouse gas radiative forcing in both the full-forcing ensembles (Fig. 1b and d), whereas the single forcing simulations do not show this 20th century warming as they do not contain the greenhouse-gas radiative forcing.

3 Model – (pseudo-proxy) data comparison setup

A pseudo-proxy series can be defined as an instrumental or climate model data series that has purposefully been distorted through the addition of noise (Jones et al., 2009; Smerdon, 2012). This is to ensure that the pseudo-proxies account for a fraction of the variance of a temperature series, as is the case for a real proxy reconstruction of temperature. A key advantage of this approach is that the distortion and reconstruction targets are both prescribed and hence fully known. Here, the pseudo-proxy setup is described in relation to the statistical framework, upon which further details can be read in Part 1. In the present pseudo-proxy analysis, the true temperature τ_i is defined explicitly by a particular simulation, chosen either from the E1 or E2 full-forcing ensembles, where the regions used in the comparison are specified. Then the proxy series z_i and instrumental series y_i can be constructed as τ_i plus added noise at specified levels.

An additional advantage of the pseudo-proxy approach using model output is that the number of locations can be varied from a single grid box to any number of locations. We also consider an average single time series for the entire globe. Given a realistic amount of noise in the pseudo-proxies, it is hoped, first, that the correlation-based test statistic U_R will indicate that a forced simulation from either the E1 or E2 ensemble is able to explain some of the simulated variability in another simulation from E1 or E2, when a single member of one of those forced ensembles is used as the "truth". Then, if this happens, it is hoped that the distance-based performance metric U_T will distinguish the E1 and E2 ensemble simulations from CTRL simulations, and also correctly rank them against each other, again when a single member of one of the two forced ensembles is used as the "truth". If this is not the case, then the method cannot be expected to help better constrain the definition of a suitable past millennial solar forcing amplitude, if the analysis were applied to real proxy and instrumental data.

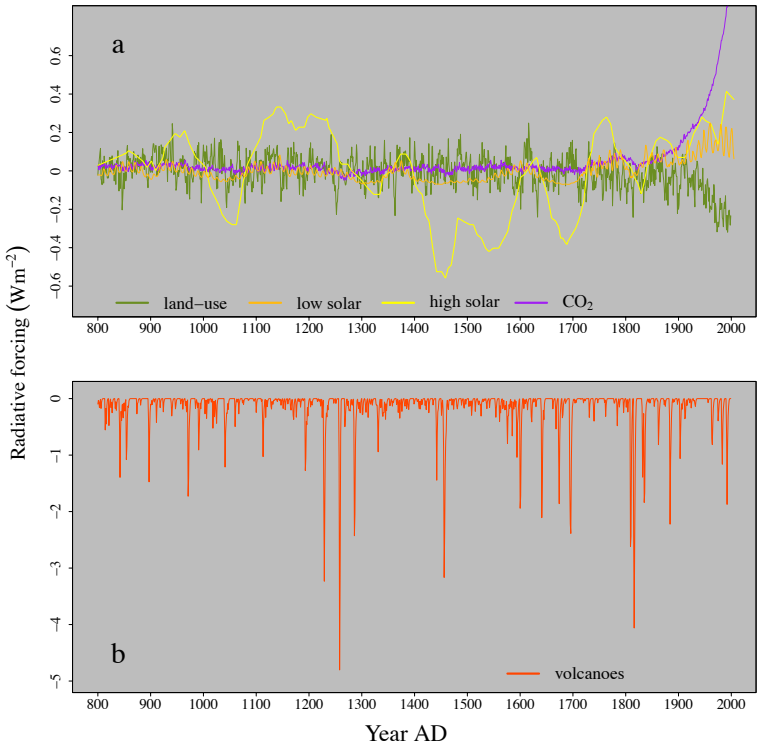

Fig. 2. Annual mean radiative forcing at the top of the atmosphere (Wm^{-2}) for **(a)** low solar (light orange), high solar (yellow), CO_2 (purple) and land-cover change (green); and for **(b)** volcanoes (red).

In our experiment, we also compare simulated temperatures from the single-forcing simulations with pseudo-proxy temperatures created from either one of the E1 or E2 ensembles, to learn more about the detectability of the effect of single forcings and their influence on temperatures in a full-forced "noisy proxy world". In all cases, the climate model simulation time sequences x_i are 2-m (surface) temperatures from the COSMOS simulations (land points only), where the forced component $\alpha\xi_i$ is the response to either a single forcing in the case of land-use changes, solar and volcanic, or to the combined forcings in the E1/E2 ensembles. Note that $\alpha = 0$ in the case of the unforced CTRL simulations (see Statistical Models 1 and 2 in Part 1).

We undertook our analysis using 30-yr non-overlapping means of simulated temperatures from the COSMOS simulations. A motivation for this choice is given later in this section. The instrumental measurements y_i are defined as the target simulation (i.e. one member from E1 or E2) for a given location over the period 1850–2000 with added white noise (θ_i), defined as representing 10 % of the total variance of y. Regarding the added noise in y_i, this approximately corresponds to a doubling of recent single-thermometer measurement error estimates (Folland et al., 2001; Brohan et al., 2006), but is chosen here on an ad hoc basis to provide a level of noise that is not negligible but yet notably smaller than in most real proxy data. The proxy series z_i are defined similarly, though over the period 1000–2000 and feature added white noise (ϵ_i) with two-thirds the total variance of z. This corresponds to an SNR = 0.71 (signal-to-noise ratio; see Smerdon, 2012) and correlation $r = 0.58$ between z and τ, which is not untypical for high-quality real proxy records (Christiansen and Ljungqvist, 2011, 2012). To represent both better and worse real proxies, considerably higher and lower percentages (always defined for the 30-yr time unit) of noise levels were also investigated (see Supplement).

The analysis included data for the period 1000–2000 AD, despite that forced simulations begin at 850 AD. The computation of the test statistics U_T (Eq. 18; Part 1) and U_R (Eq. 23; Part 1), however, was restricted to the period 1000–1850 to avoid the influence of anthropogenic greenhouse gas increases. It should also be noted that data after 1850 were used for the calibration of z_i against y_i and for estimating the total variance of y. The statistical framework of Part 1 allows for uncertainty in both the instrumental and proxy series, which are specified through a time-dependent weighting w_i (Eq. 9-12 in Part 1). In our experiment, however, the precision of z_i does not vary with time. The variance of the "true" unforced temperature, s_η^2, was estimated using detrended pseudo-instrumental data, whilst the sample variance of internal unforced variability, s_δ^2, was estimated from CTRL simulations (see Sect. 5 in Part 1).

As described in Sect. 2 of Part 1, the unforced simulated temperature δ_i is assumed to be white noise. It is of course quite possible that white noise is not a good representation

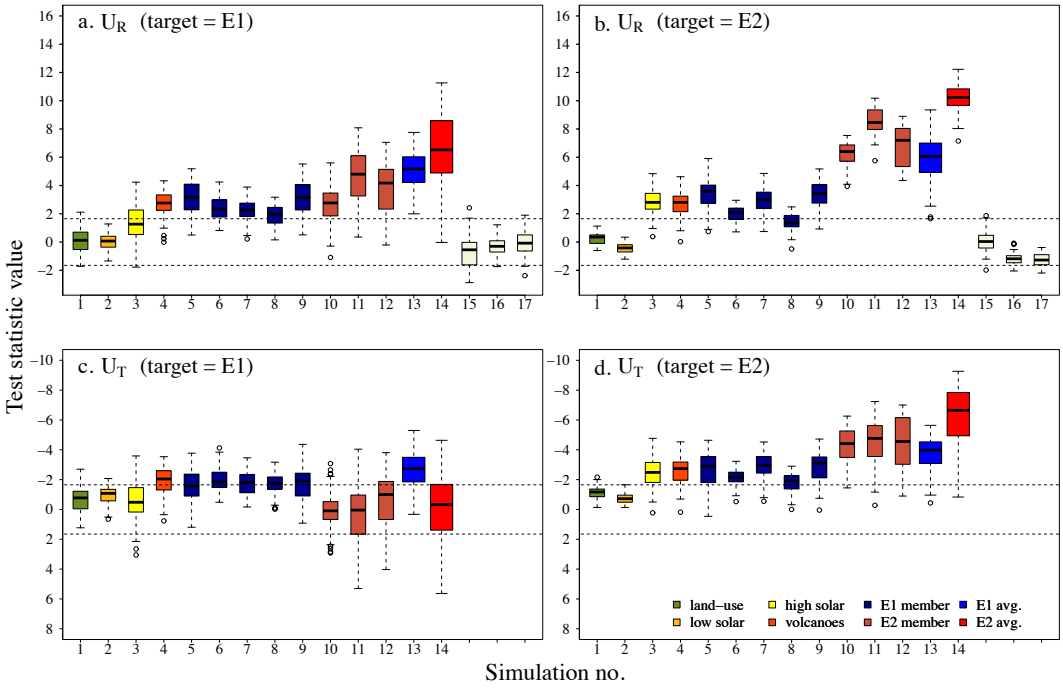

Fig. 3. Box plots for U_R correlation (top panel) and U_T distance (bottom panel) test statistics for each of the global land-only average COSMOS simulation temperature series, compared to ≈100 different pseudo-proxy temperature realizations (iteratively running through the E1 and E2 ensemble members as targets – see text). The left panels are for E1 ("low" solar) as target, right E2 ("high" solar) as target. The 5 % two-sided significance levels are shown with dashed lines. Each box covers the 50 % interval between the lower and upper quartiles, with the median as a thick black line between. The simulations are: 1 = land-use changes, 2 = low solar, 3 = high solar, 4 = volcanoes, 5–9 = E1, 10–12 = E2, 13 = average E1, 14 = average E2. The CTRL simulation (numbers 15–17) results are shown for the U_R analysis but not for U_T, since they are then used as internal references. Note that the y-axis for U_T is flipped to simplify any comparisons with the U_R box plots.

of the internal variability of the true climate, and the distance measure D^2 does not require white noise. However, the null hypothesis of the statistical tests is that forced simulations are equivalent to CTRL simulations, so for the described tests to have the prescribed type I error level, the unforced simulations should be well represented by white noise. We investigated the seriousness of this problem by calculating the lag-1 autocorrelation for the full 3000-yr CTRL simulation, both in terms of the proportion of global area with significant autocorrelations for various time resolutions, as well as the lag-1 autocorrelation for the global land-only series (see Supplement for further details). It was found that beyond a 20-yr time resolution, δ_i can be considered as white noise, in keeping with the statistical assumptions of Sect. 2 of Part 1. Hence, a non-overlapping 30-yr mean resolution, as used in the present analysis, should be able to keep the type I error of the tests under reasonable control in the model.

4 Model – (pseudo-proxy) data comparison

4.1 Global analysis

We first conducted a study on globally averaged (area-weighted) time series using only land points (i.e. the data shown in Fig. 1), the results of which are shown in Fig. 3 (Fig. 3a and b show the U_R correlation analysis results, whilst Fig. 3c and d show the U_T distance measure results). To clarify, a positive U_R represents a positive correlation between a simulation and its target, whilst a negative U_T indicates a better performance of a forced simulation compared with unforced simulations. The global mean was investigated first, simply because this series will likely exhibit a stronger signal-to-noise ratio of the forced component than at the individual grid-point scale where internal temperature variability is more dominant (Servonnat et al., 2010). Hence, we use a single set of τ_i, y_i and z_i sequences in this globally averaged analysis (i.e. the summation in the definitions of U_T and U_R is made over a single term, and no covariance computations are needed). Both the E1 and E2 simulations were used separately as targets in this experiment, and to use as many target

"true" climates or "truths" as possible, each ensemble member was used as the target in turn.

For each type of "truth", ≈ 100 noise realizations were generated to produce y_i and z_i with a rotation in the five E1 target simulations (20 noise realizations for each simulation, $5 \times 20 = 100$) (Fig. 3a and c) and in the three E2 target simulations (33 noise realizations for each simulation, $3 \times 33 = 99$) (Fig. 3b and d). Iteratively treating the E1 or E2 ensemble members as targets could cause the distributions to be hierarchical, in that the error distribution associated with different noise realizations could potentially be small in comparison with the difference between ensemble members (internal climate variability in the model). Hence, an identical analysis to this was conducted but with zero proxy noise added to the target temperatures, which revealed the E1 and E2 ensemble simulations to give results with little qualitative spread (*not shown*). This satisfied the authors sufficiently that the *spread* of the distributions in Fig. 3 predominantly represents the uncertainty due to the pseudo-proxy noise realizations.

To further explain the U_T and U_R box plot distributions shown in Fig. 3, the first four represent the single forcing simulations, namely land-use changes (green), low solar (light orange), high solar (yellow) and volcanoes (red), where they are compared with either the E1 (left panels) or the E2 (right panels) simulations as target. Analogously, the next five box plots (numbers 5–9) represent the E1 simulations, all coloured dark blue with their corresponding ensemble average U_R/U_T value in blue (number 13). The three E2 simulations are coloured dark red (numbers 10–12) with their corresponding ensemble average in red (number 14). Note that, when an E1 (or E2) simulation is used as the target, this target simulation is excluded from the E1 (or E2) ensemble being analysed. Additionally, for comparison, Fig. 3a and b feature an analysis of the three CTRL simulation segments (numbers 15–17) as these are not required in the calculation of U_R.

From Fig. 3a and b, the U_R correlation analysis, it is clear that individual E1 and E2 ensemble members are significantly correlated with *both* E1 and E2 targets. However, the E2 simulations are the most highly correlated, whichever is the target. This can be expected in so far as the E2 simulations feature the strongest solar forcing and the largest variability (Fig. 1). However, the significant correlations between E1 and E2 ensembles may not be reflected in a distance-based measure. U_T is expected to be more effective in distinguishing between the simulations and, in some instances, being capable of ranking them. The principal reason being that the correlation analysis does not consider the variance of two compared series (target and simulation), whereas this is explicitly considered in the distance measure. This can be seen by the fact that, when E1 serves as target (Fig. 3c), E1 simulations are generally significantly closer to the target than CTRL simulations, whilst the E2 simulations are not. The E1 and E2 simulations are also correctly distinguished when E2 serves as target. In this case, however, both are closer to the target than CTRL simulations (Fig. 3d).

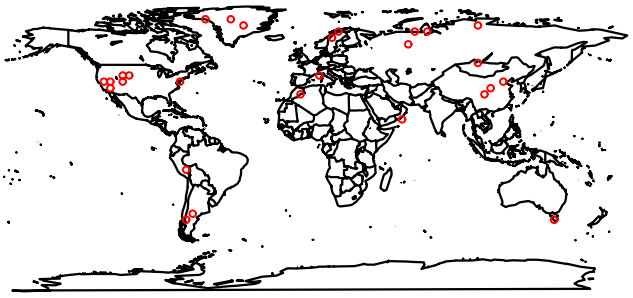

Fig. 4. The 27 proxy locations taken from Juckes et al. (2007) for the present local-scale comparison. Note that the Juckes et al. (2007) set consists of 33 proxy locations, but some locations were so close together that a single representation was chosen for that location. A higher resolution model would likely have allowed a comparison using the full set of locations.

The low solar single-forced simulation (number 2) is not significantly correlated with, or close to, the E1 targets (Fig. 3a and c). In contrast, the high solar simulation (number 3) is significantly correlated with, and close to, the E2 targets (Fig. 3b and d). This implies that the low solar forcing is too weak to produce any detectable effect at the 30-yr time unit, whilst the high solar *is* strong enough. A related conclusion was reached by Ammann et al. (2007): the greater the solar forcing amplitude applied to their model, the weaker the detectable response to other natural forcings. In regard to CTRL simulations, their U_R values are mostly insignificant, as should be expected given the construction of the experiment and the null hypothesis being tested.

4.2 Local analysis

At global or hemispheric scales, the temperature can be expected to respond to large-scale external forcings (such as solar or greenhouse gases), whereas at local or regional scales the internal climate dynamics can account for a larger proportion of the temperature variability (Goosse et al., 2005). Hence, on small spatial scales, the ability to distinguish between simulations that use low and high solar forcing, and consequently rank them, may not be possible. A current set of proxy locations from Juckes et al. (2007) was used to generate pseudo-proxies in order to investigate whether the low and high solar simulations can still be distinguished (Fig. 4). Though this set of locations is clearly a sparse representation of the global surface, 20–40 or so proxy locations is a typical number of high quality millennial proxy data found in current analyses (Christiansen and Ljungqvist, 2011).

The same type of experiments conducted in the global analysis (Fig. 3) was also conducted for the combined Juckes et al. (2007) locations (Fig. 5). Specifically, we compute local correlation and distance measures for each proxy location,

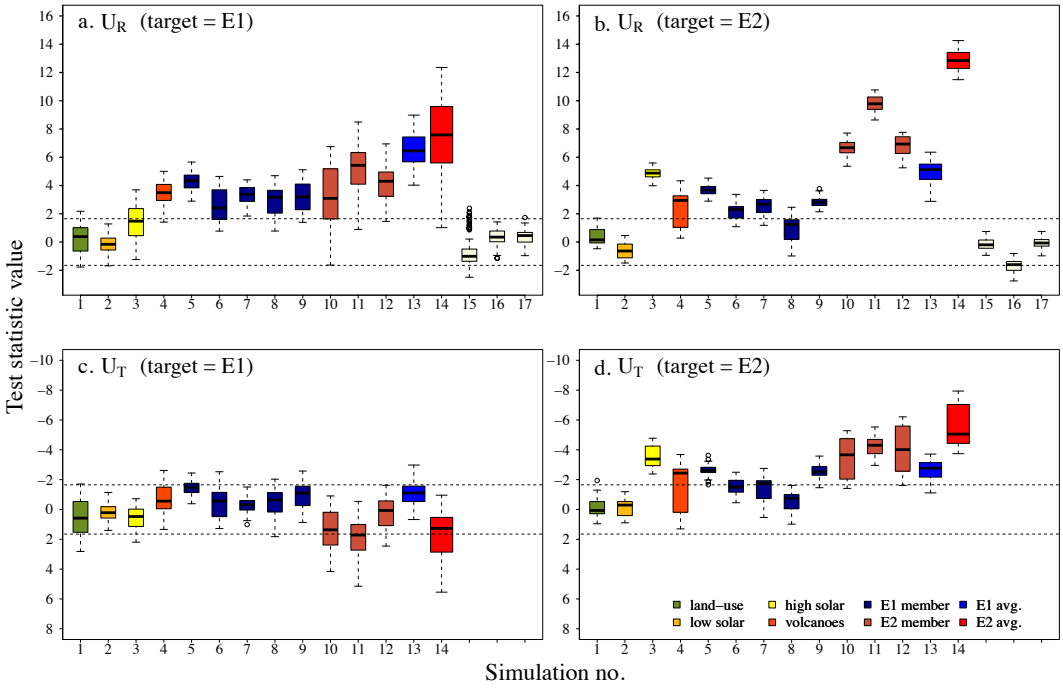

Fig. 5. As Fig. 3, but using the local proxy locations from Juckes et al. (2007).

before they are combined to obtain a single U_R and U_T value for each simulation (Sects. 7 and 8; Part 1). The correlation analysis U_R for the Juckes et al. (2007) proxy locations (Fig. 5a and b) gives similar results to the global time-series analysis, though surprisingly the correlations are not less significant, rather sometimes even more significant. This is something that could not have been expected due to the increased influence of internal (unforced) variability at the regional scale in combination with the reduced area coverage. However, in contrast to the global analysis, when E1 serves as target, U_T is unable to distinguish the E1 simulations from the CTRL simulations (Fig. 5c), whereas the E2 simulations are again significantly closer to the target than the CTRL simulations when E2 serves as target (Fig. 5d). Concerning the single forcing simulations, only the high solar (number 3) is significantly closer to the targets than the CTRL simulations when E2 is the target (Fig. 5d).

Using a realistic set of proxy locations such as the Juckes et al. (2007) set, it seems difficult to rank simulations, unless the forcing is large and multi-decadal in nature (as is the case for the high solar forcing used here). Note that U_R is more sensitive than U_T for testing if a model forcing has any correspondence with the true climate, but it answers a different question than U_T. This higher sensitivity is seen when we compare subfigures a and b with c and d respectively in both Figs. 3 or 5. Specifically, if U_R is not significant, nor is U_T. Comparisons between the Juckes et al. (2007) and global land-only average results naturally lead to the question of how the possibility to rank simulations depends on the spatial coverage of the pseudo-proxy data.

5 Varying coverage

There are in practice relatively few locations which have high quality proxy data available or where there is the potential at present to acquire more data. A pseudo-proxy experiment, however, has the advantage of allowing any number of locations to be used to serve as a proxy series or instrumental series. Hence, an analysis is conducted on how varying degrees of % surface area coverage affect the sensitivity of the correlation and distance measures to distinguish between simulations with low or high solar forcing.

The various specified global surface area coverages are for 0.1, 0.25, 0.5, 1, 2, 3, 4, 5 %, using only land grid points, which is equivalent to 3, 10, 22, 44, 90, 137, 183, 230 proxy locations. Calculation of the covariance matrices $\mathrm{Cov}(T_{j_1}, T_{j_2})$ (Sect. 7; Part 1) and $\mathrm{Cov}(R_{j_1}, R_{j_2})$ (Sect. 8; Part 1) becomes computationally intensive for large % coverages; hence, they were only calculated up to 5 %. Note that, although a principal component truncation could in principle be considered here to reduce the dimensionality of the climate variability represented by the proxy series, it was felt that, due to the heterogeneous coverage distributions and the arbitrary nature of the choice of retained principal components (and also considering the varying seasonal representation and time periods covered by *real* proxies), we would not conduct such an approach here. The set of proxy locations

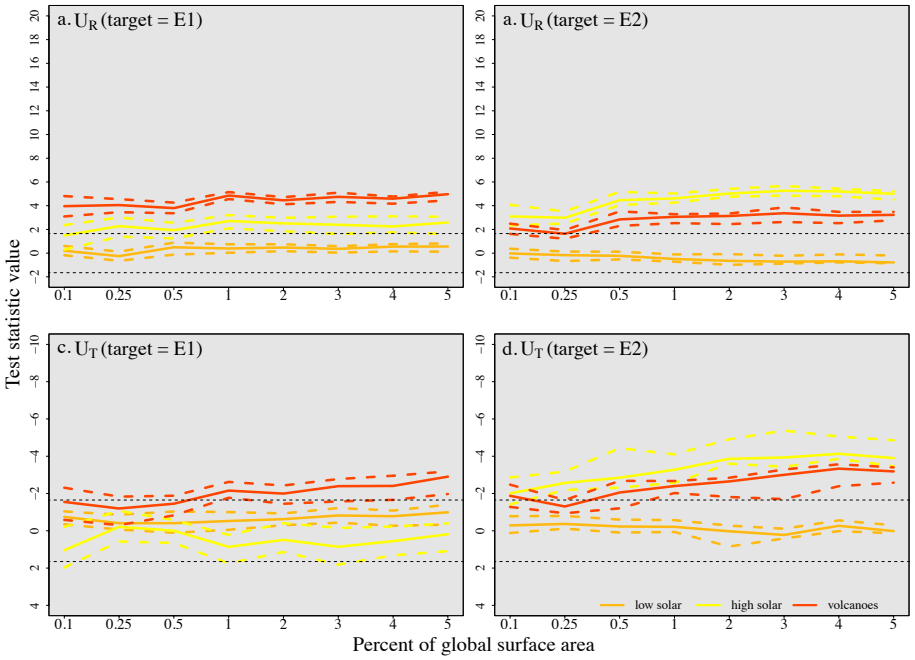

Fig. 6. U_R correlation (top panels) and U_T distance (bottom panels) measures for volcanic (red), low (light orange) and high (yellow) solar forcing simulations, against increasing % global surface area coverage. The left panels are for E1 as target, right panels E2 as target. The 5 % significance level is shown with dashed lines. The filled coloured lines denote the median value, with the dashed coloured lines representing the upper and lower quartiles.

was selected as a stratified random sample from the available land points in the COSMOS simulations, with specified proportions for three strata (the latitudinal bands 0–30°, 30–60°, 60–90°). The stratification was chosen to better control the coverage and to account for the changing area of the grid points with latitude in the simulations.

Figure 6 shows the correlation U_R (top panel) and distance U_T (bottom panel) measures for the low (light orange) and high solar (yellow) single forcing simulations for different % coverages, again with both E1 and E2 simulations serving as targets. For each % coverage level, approximately 100 noise realizations were generated, of which the median values are represented by solid lines and the upper and lower quartiles are dashed. For comparison, results for the volcanic (red) forcing simulation are also shown. The target and test statistics panels are arranged the same as Figs. 3 and 5.

The high solar simulation is significantly correlated even for the lowest coverages when E2 serves as target (Fig. 6b), whilst also achieving significant U_R values for coverages upwards of 1 % when E1 serves as target. Contrastingly, the high solar simulation U_T values are significantly better than the CTRL simulations for all coverages when E2 serves as target (Fig. 6d), but not for any coverage when E1 serves as target (Fig. 6c). The low solar simulation shows no significant correlations for either target ensemble and can therefore be expected to be indistinguishable from the CTRL simulations using the U_T measure. The volcanic simulation is mostly significantly correlated with both E1 and E2 targets (Fig. 6a and b), but its U_T values are generally only significant for coverages upwards of 1 % for both targets (Fig. 6c and d).

Figure 7 is arranged as Fig. 6 but shows the E1 (blue) and E2 (red) ensemble average results. Both ensembles are significantly correlated with all targets, even for the lowest data coverages. The results for U_T are much the same as for the global analysis in Sect. 4.1, where the E1 and E2 ensembles can be correctly ranked with their respective targets. For coverages lower than 1 %, it becomes difficult to distinguish E1 from the CTRL simulations or separate the E1 and E2 simulations when E1 serves as target (Fig. 7c). Additionally, the experiments of Figs. 6 and 7 were conducted for cases with a SNR = 0.25 and also with negligible noise, the results of which are briefly discussed in the conclusions and shown in the Supplement. An important feature of note in Figs. 6 and 7 is how flat the U_R and U_T measures are with changing % coverage after a certain coverage is reached. In fact, there is little gain in increasing the sample size from 40 or so proxy series to several hundred. Above all else, this suggests a substantial degree of spatial correlation in simulated temperatures, given the 30-yr time resolution used in this analysis (Jones et al., 1997; Franke et al., 2011).

Finally, we should mention that two variants of U_T for ensemble means were defined in Part 1. In the main variant, averaging of a distance measure D^2 is undertaken for

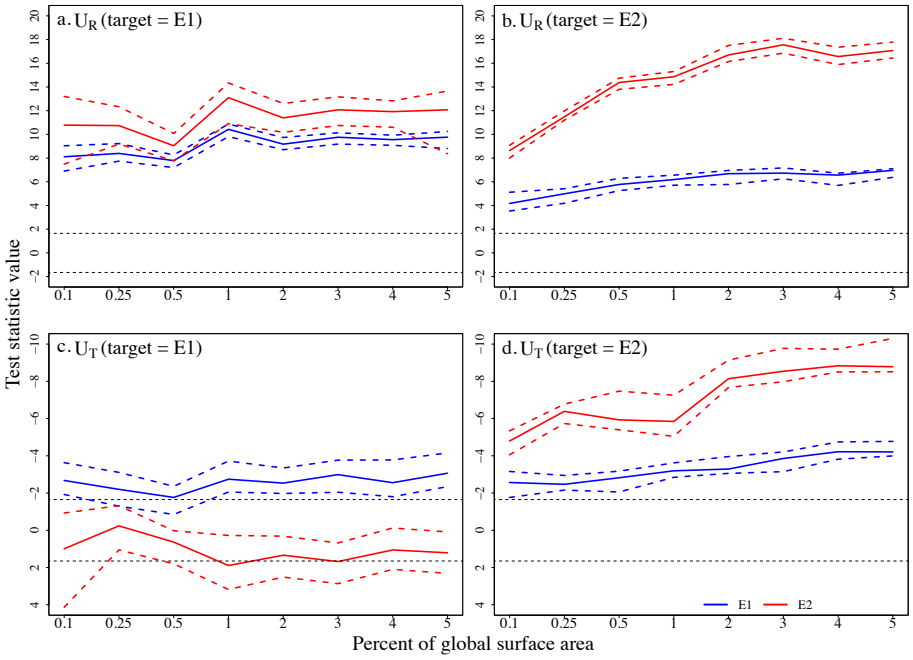

Fig. 7. As Fig. 6, but for the E1 (blue) and E2 (red) ensemble averages.

different individual simulations before calculating the T and U_T statistics. This variant has been used in all analyses here. In the alternative variant, defined in Appendix A of Part 1, the averaging is instead undertaken on the simulation temperature series before calculating the D^2. Results for varying coverage with this alternative approach are shown in Appendix A of this paper, where the Fig. A1 should be compared with Fig. 7c and d.

6 Conclusions

We apply a new statistical framework (Sundberg et al., 2012) designed for comparing ensemble model simulation surface temperature from one or more locations with proxy and instrumental data. This framework derives a unified correlation-based statistic (U_R) that provides an initial test of whether a set of simulation time series from different locations (and/or seasons) correlates with a set of target series for the corresponding real locations (seasons), and a distance-based measure (U_T) that can be used to assess the goodness-of-fit of a given forced simulation in comparison with those that are unforced. The ultimate goal was to rank the simulations according to their closeness to the target data. A pseudo-proxy experiment was designed for this task, based on the MPI-COSMOS Earth system model simulations (Jungclaus et al., 2010). Here, the "true" climate and the proxy noise are known; hence, if no difference between two forced simulations containing different solar forcing evolutions can be detected with these methods for realistic proxy noise levels, then no significant conclusions could be assumed based on comparing the same model output with real proxy data.

Firstly, an analysis was conducted on globally averaged land-only data where a single series was calculated for each simulation and compared with every member of the full-forced E1 (with low solar) and E2 (with high solar) ensembles in turn plus added noise. Regardless of whether E1 or E2 simulations are used as a target, it was found that both simulation types are strongly correlated (significant positive U_R) with each other. Knowing that the shared forcing information gives significantly correlated temperature evolutions between both low and high solar simulations, U_T was found capable of ranking these simulations correctly.

Given that this statistical framework has been developed in view of using real proxy information to assess the goodness-of-fit of model simulations, a pseudo-proxy evaluation was also conducted for a representative set of about 30 proxy locations (taken from Juckes et al., 2007). The results of this multiple-site local comparison were similar to the global land-only results; however, the U_T values of the E1 ensemble could not be said to be significantly different from the CTRL simulations when E1 serves as target. This motivated an analysis of how differing % coverage levels change the significance of the U_T and U_R statistics (Figs. 6 and 7). The results suggest that, for a global coverage of say 40 or more proxy locations, if a high quality of individual proxy series is obtained with low noise levels (SNR of at least 0.71 for white noise defined at the analysed 30-yr time unit), it can be possible to distinguish the E1 and E2 ensembles when E1 serves as target. If E2 serves as target, very few proxy

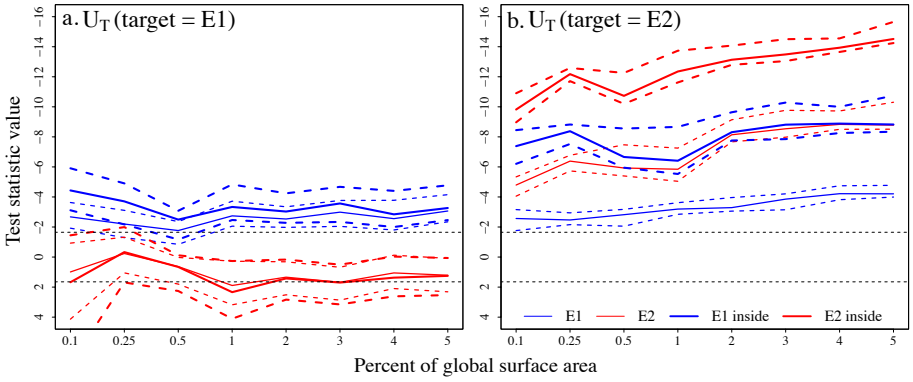

Fig. A1. As Fig. 7, with E1 (blue) and E2 (red) ensemble averages. The thick lines denote the use of inside averaging, whilst the thin lines denote outside averaging (as presented in Fig. 7). Note that the y-axis is extended here to accommodate the inside averaging lines.

series are needed. Additionally, the same type of analysis was conducted for a higher noise level (SNR = 0.25), where it was found that the E1 and E2 simulations are indistinguishable even if the global surface area coverage is 5 % (approximately 230 proxy locations). Although these results, in quantitative terms, are conditional upon the actual climate model simulations used to define the pseudo-proxy world, they have an important implication: it is more important to improve the quality of individual local proxy series in terms of SNR than it is to increase the quantity of available proxy locations. Even a limited spatial coverage is sufficient to distinguish forced multi-decadal temperature signals, provided the temperature proxies are of a sufficient quality and represent areas that can be directly compared with model output.

Appendix A

Averaging inside D^2

From Fig. A1, if the alternative "inside" averaging (defined in Appendix A of Part 1; thick lines) is used instead of "outside" averaging (thin lines) in calculating the E1 and E2 ensemble averages, the U_T results appear to change little if E1 serves as target, whereas there is a substantial increase in the significance of U_T when E2 serves as target. This likely reflects the fact that, if there is a stronger common signal amongst the ensemble members (as with the high solar E2 ensemble), then the inside averaging approach will enhance the SNR of the series, whilst, if the common signal is weaker (as with the low solar E1 ensemble), there will not be a large difference between the approaches. Hence, inside averaging can be more effective than outside averaging.

Supplementary material related to this article is available online at: http://www.clim-past.net/8/1355/2012/cp-8-1355-2012-supplement.pdf.

Acknowledgements. This research was funded by the Swedish Research Council (grants 70454201, 90751501 and B0334901) and the European Union (FP6 grant 017008, "Millennium" project). We thank Johann Jungclaus of the Max Planck Institute for providing the COSMOS data as well as help and advice regarding the simulations. We also thank G. Hegerl, Y. H. Yamazaki and an anonymous reviewer for constructive comments and advice in their reviews of the discussion paper.

Edited by: P. Brohan

References

Ammann, C. M., Joos, F., Schimel, D. S., Otto-Bliesner, B. L., and Tomas, R. A.: Solar influence on climate during the past millennium: Results from transient simulations with the NCAR Climate System Model, P. Natl. Acad. Sci., 104, 3713–3718, 2007.

Bard, E., Raisbeck, G., Yiou, F., and Jouzel, J.: Solar irradiance during the last 1200 years based on cosmogenic nuclides, Tellus B, 52, 985–992, 2000.

Brohan, P., Kennedy, J. J., Harris, I., Tett, S. F. B., and Jones, P. D.: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850, J. Geophys. Res., 111, D12106, doi:10.1029/2005JD006548, 1–12, 2006.

Christiansen, B. and Ljungqvist, F. C.: Reconstruction of the extra-tropical NH mean temperature over the last millennium with a method that preserves low-frequency variability, J. Climate, 24, 6013–6034, 2011.

Christiansen, B. and Ljungqvist, F. C.: The extra-tropical Northern Hemisphere temperature in the last two millennia: reconstructions of low-frequency variability, Clim. Past, 8, 765–786, doi:10.5194/cp-8-765-2012, 2012.

Cliver, E. W., Boriakoff, V., and Feynman, J.: Solar variability and climate change: a geomagnetic and aa index and global surface temperature, Geophys. Res. Lett., 25, 1035–1038, 1998.

Feulner, G.: Are the most recent estimates for Maunder Minimum solar irradiance in agreement with temperature reconstructions?, Geophys. Res. Lett., 38, L16706, doi:10.1029/2011GL048529, 2011.

Folland, C. K., Rayner, N. A., Brown, S. J., Smith, T. M., Shen, S. S. P., Parker, D. E., Macadam, I., Jones, P. D., Jones, R. N., Nicholls, N., and Sexton, D. M. H.: Global temperature change and its uncertainties since 1861, Geophys. Res. Lett., 28, 2621–2624, 2001.

Franke, J., Gonzalez-Rouco, J. F., Frank, D., and Graham, N. E.: 200 years of European temperature variability: insights from and tests of the proxy surrogate reconstruction analog method, Clim. Dynam., 37, 133–150, 2011.

Goosse, H., Renssen, H., and Bradley, R. S.: Internal and forced climate variability during the last millennium: a model-data comparison using ensemble simulations, Quaternary Sci. Rev., 24, 1345–1360, 2005.

Gray, L. J., Beer, J., Geller, M., Haigh, J. D., Lockwood, M., Matthes, K., Cubasch, U., Fleitmann, D., Harrison, G., Hood, L., Luterbacher, J., Meehl, G. A., Shindell, D., van Geel, B., and White, W.: Solar influences on climate, Rev. Geophys., 48, RG4001, doi:10.1029/2009RG000282, 2010.

Hoyt, D. V. and Schatten, K. H.: A discussion of plausible solar irradiance variations, J. Geophys. Res., 98, 18895–18906, 1993.

Jones, P. D., Osborn, T. J., and Briffa, K. R.: Estimating Sampling Errors in Large-Scale Temperature Averages, P. Natl. Acad. Sci., 10, 2548–2568, 1997.

Jones, P. D., Briffa, K. R., Osborn, T. J., Lough, J. M., van Ommen, T. D., Vinther, B. M., Luterbacher, J. W. E. R. Z. F. W., Mann, M. E., Schmidt, G. A., Ammann, C. M., Buckley, B. M., Cobb, K. M., Esper, J., Goosse, H., Graham, N., Janse, E., Kiefer, T., Kull, C., Küttel, M., Mosley-Thompson, E., Overpeck, J. T., Riedwyl, N., Schulz, M., Tudhope, A. W., Villalba, R., Wanner, H., Wolff, E., and Xoplaki, E.: High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects, Holocene, 19, 3–49, 2009.

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., 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.

Knutti, R. and Hegerl, G. C.: The equilibrium sensitivity of the Earth's temperature to radiation changes, Nat. Geosci., 1, 735–743, 2008.

Krivova, N. A., Balmaceda, L., and Solanki, S. K.: Reconstruction of solar total irradiance since 1700 from the surface magnetic flux, Astron. Astrophys., 467, 335–346, 2007.

Lean, J., Beer, J., and Bradley, R.: Reconstruction of solar irradiance since 1610: Implications for climate change, Geophys. Res. Lett., 22, 3195–3198, 1995.

Lockwood, M.: Shining a light on solar impacts, Nat. Clim. Change, 1, 98–99, 2011.

Marsland, S. J., Haak, H., Jungclaus, J. H., Latif, M., and Roeske, F.: THe Max Planck Institute global ocean/ice model with orthogonal curvilinear coordinates, Ocean Modell., 5, 91–127, 2003.

Reid, G. C.: Solar total irradiance variations and the global sea surface temperature record, J. Geophys. Res., 96, 2835–2844, 1991.

Reid, G. C.: Solar forcing and the global climate change since the mid-17th century, Climate Change, 37, 391-405, 1997.

Roeckner, E., Bäuml, G., Bonaventura, L., Brokopf, R., Esch, M. G., Hagemann, S., Kirchner, I., Kornblueh, L., Manzini, E., Rhodin, A., Schlese, U., Schulzweida, U., and Tompkins, A.: The atmospheric general circulation model ECHAM5, Part I: Model description, Technical Report, Max Planck Institute of Meteorology, 349, available from MPI for Meteorology, Hamburg, Germany, 127 pp., 2003.

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.

Servonnat, J., Yiou, P., Khodri, M., Swingedouw, D., and Denvil, S.: Influence of solar variability, CO_2 and orbital forcing between 1000 and 1850 AD in the IPSLCM4 model, Clim. Past, 6, 445–460, doi:10.5194/cp-6-445-2010, 2010.

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, 1–8, 2011.

Smerdon, J. E.: Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments, WIREs Clim. Change, 3, 63–77, doi:10.1002/wcc.149, 2012.

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.

Tapping, K. F., Boteler, D., Charbonneau, P., Crouch, A., Manson, A., and Paquette, H.: Solar magnetic activity and total irradiance since the Maunder Minimum, Solar Phys., 246, 309–326, 2009.

Wang, Y.-M., Lean, J. L., and Sheeley, N. R.: Modeling the Sun's magnetic field and irradiance since 1713, Astrophys. J., 625, 522–538, 2005.

Yoshimori, M., Stocker, T. F., Raible, C. C., and Renold, M: Externally forced and internal variability in ensemble climate simulations of the Maunder Minimum, J. Climate, 18, 4253–4270, 2005.

Zhang, Q., Soon, W. H., Baliunas, S. L., Lockwood, G. W., Skiff, B. A., and Radick, R. R.: A method of determining possible brightness variations of the sun in past centuries from observations of solar-type stars, Astrophys. J., 427, L111–L114, 1994.