

Past millennial solar forcing magnitude

A statistical hemispheric-scale climate model versus proxy data comparison

Alistair Hind · Anders Moberg

Received: 20 July 2012 / Accepted: 6 September 2012 / Published online: 22 September 2012
© The Author(s) 2012. This article is published with open access at Springerlink.com

Abstract A set of global climate model simulations for the last thousand years developed by the Max Planck Institute is compared with paleoclimate proxy data and instrumental data, focusing on surface temperatures for land areas between 30° and 75°N. The proxy data are obtained from six previously published Northern Hemispheric-scale temperature reconstructions, here recalibrated for consistency, which are compared with the simulations utilizing a newly developed statistical framework for ranking several competing simulations by means of their statistical distance against past climate variations. The climate model simulations are driven by either “low” or “high” solar forcing amplitudes (0.1 and 0.25 % smaller total solar irradiance in the Maunder Minimum period compared to the present) in addition to several other known climate forcings of importance. Our results indicate that the high solar forcing amplitude results in a poorer match with the hemispheric-scale temperature reconstructions and lends stronger statistical support for the low-amplitude solar forcing. However, results are likely conditional upon the sensitivity of the climate model used and strongly dependent on the choice of temperature reconstruction, hence a greater consensus is needed regarding the reconstruction of past temperatures as this currently provides a great source of uncertainty.

Keywords Climate model evaluation · Climate proxy data · Solar forcing · Last millennium · Northern Hemisphere

Electronic supplementary material The online version of this article (doi:10.1007/s00382-012-1526-6) contains supplementary material, which is available to authorized users.

A. Hind (✉) · A. Moberg
Department of Physical Geography and Quaternary Geology,
Bert Bolin Centre for Climate Research, Stockholm University,
106 91 Stockholm, Sweden
e-mail: alistair.hind@natgeo.su.se

1 Introduction

Solar irradiance can impact upon the global climate through its variations (Gray et al. 2010). If changes in the shape of the emitted solar spectrum are neglected, these variations can be quantified as the total solar irradiance (TSI). TSI has been directly measured using instruments mounted on satellites as of the mid-1970s, however before this period it must be reconstructed based on proxy indices of solar irradiance found to correlate with TSI variations during the instrumental period. Recent estimates of these solar activity indices have generally been based on sunspot numbers and/or solar magnetic flux changes (derived from geomagnetic information or cosmogenic isotopes). Although it has been established that TSI varies in phase with the 11-year sunspot cycle with an amplitude of about 0.1 %, there exists a great deal of debate surrounding the magnitude of centennial time-scale TSI variations—or indeed if such variations exist at all (Lockwood 2011). In climate science, the magnitude of TSI variations is often portrayed as the hypothesized reduction, compared to present values, during the Maunder Minimum period (AD 1645–1715; characterized by low solar activity Eddy (1976). This reduction is widely regarded, of present, as being in the range 0.04–0.1 %, as adopted by the Paleoclimate Model Intercomparison Project Phase III (PMIP3) (Schmidt et al. 2011). Two very recent publications have reignited debate over the magnitude of background TSI variations however. Schrijver et al. (2011) argue that long-term TSI variations may be even weaker than the smallest used in PMIP3 (potentially negligible even), whilst Shapiro et al. (2011) suggest a decrease during the Maunder Minimum of more than 0.4 %. This, in turn, would correspond in magnitude to about 40 % of the radiative forcing from increased greenhouse-gas concentrations over the twentieth century (Lockwood 2011).

Given this uncertainty in long-term background TSI variations, it seems instructive to compare alternative forcing histories as drivers of climate model simulations. These simulations are hoped to be useful in describing a climate outside of the instrumental period (Schmidt 2010; Schmidt et al. 2011) and the temperature output from these models can then be compared with observed temperatures and past temperatures reconstructed from proxy data (Hegerl et al. 2007; Mann et al. 2009; Jungclaus et al. 2010; Braconnot et al. 2012). It may be argued that the simulation most similar to the instrumental or reconstructed temperatures is more likely to contain the correct forcing history (Feulner 2011). In practice, however, there are several factors that render such an attempted constraint of TSI variations difficult. There is a non-negligible amount of noise present in the proxy series (Jones et al. 2009), unforced variability in the climate system (Yoshimori et al. 2005) as well as uncertainty regarding the sensitivity of the Earth's climate to radiative forcing (Hegerl et al. 2006; Knutti and Hegerl 2008). In light of this, Sundberg et al. (2012) developed a statistical framework for comparing ensemble simulation surface temperature output from global climate models with instrumental and proxy temperature data, designed to address some of these difficulties while ranking alternative competing simulations by means of their estimated closeness to the unobservable true past temperature variations.

In a pseudoproxy analysis, Hind et al. (2012) used this statistical framework to compare several millennial simulations with alternative solar forcing histories with “low” and “high” amplitudes, corresponding to a 0.1 and 0.25 % reduction in TSI in the Maunder Minimum. There, it was found that for global land-only averages and realistic proxy noise levels, low and high solar forced simulations could be correctly ranked against each other, in the presence of other forcing histories and internal (model) climate variability. It is this framework that we intend to use here for the express purpose of comparing two ensemble simulations of the last millennium [the same as in Hind et al. (2012)] with published hemispheric-scale temperature reconstructions based on proxy data from the Northern Hemisphere. A motivation for the conducting of a hemispheric-scale analysis is that internal variability is likely to dominate (or be comparatively greater than) local/regional temperature anomalies, whilst on the hemispheric and global scale external forcing dominates (Servonnat et al. 2010). Additionally, there is no evidence of any global or hemispheric-scale near-surface temperature changes caused by solar-induced UV variability affecting the stratosphere, though there may be some influence on the near-surface troposphere in certain regions and seasons, such as Eurasia during the winter [see review by Lockwood (2012)]. Note that the use of the integral TSI neglects changes in spectral shape which are

largest in the UV portion. For the purposes of undertaking climate model versus proxy data comparisons, there have been many attempts at reconstructing temperature on hemispheric scales with particular emphasis on the Northern Hemisphere and the last millennium (Jansen et al. 2007; Mann et al. 2008; Frank et al. 2010; Ljungqvist 2010; Christiansen and Ljungqvist 2011). These various reconstructions involve the collection, synthesis and combination of many individual proxy series, which can vary greatly in their constituent proxies and the methods used in their combination for purposes of representing large-scale time series. To account for how these choices may affect a comparison between reconstructed temperature and equivalent simulation output, we select several Hemispheric-scale temperature reconstructions published in the last decade. The goal of this study is to investigate whether we can determine, with certainty, which of the two alternative solar forcing histories provides simulated temperatures most similar to the reconstructed temperatures, in the presence of other important climate forcings and internal climate variability.

2 Data and methods

2.1 Hemispheric-scale temperature reconstructions

Most hemispheric or global-scale temperature reconstructions employ some variant of an ‘indexing’ method, where a selection of local climate proxy records are first standardized, then combined into a single time series before being calibrated against instrumental temperatures averaged over a target region (Jones and Mann 2004; Wahl and Smerdon 2012). There are a variety of ways in which the proxy-selection process can be implemented, such as using the correlation between individual proxy records and their local temperatures (Christiansen and Ljungqvist 2011), or selection based on the assumed ability of a given record to retain low-frequency variability (Jones and Mann 2004), or simply on physically grounded expectations that the proxies have a temperature signal (Juckes et al. 2007). One may solely use data from tree-rings for example (D’Arrigo et al. 2006) or proxies from different archives may be combined (Ljungqvist 2010). Once the data have been selected, choices can also be made about the nature of the weighting in the composite averaging of series; for example no weighting (Juckes et al. 2007), weighting based on correlation with local temperatures (Hegerl et al. 2007) or a combination of local correlation and size of the area that the proxy represents (Jones and Mann 2004). Different choices have also been made regarding which large-scale region the temperature reconstruction should represent; the entire Northern Hemisphere (Jones and Mann 2004; Juckes

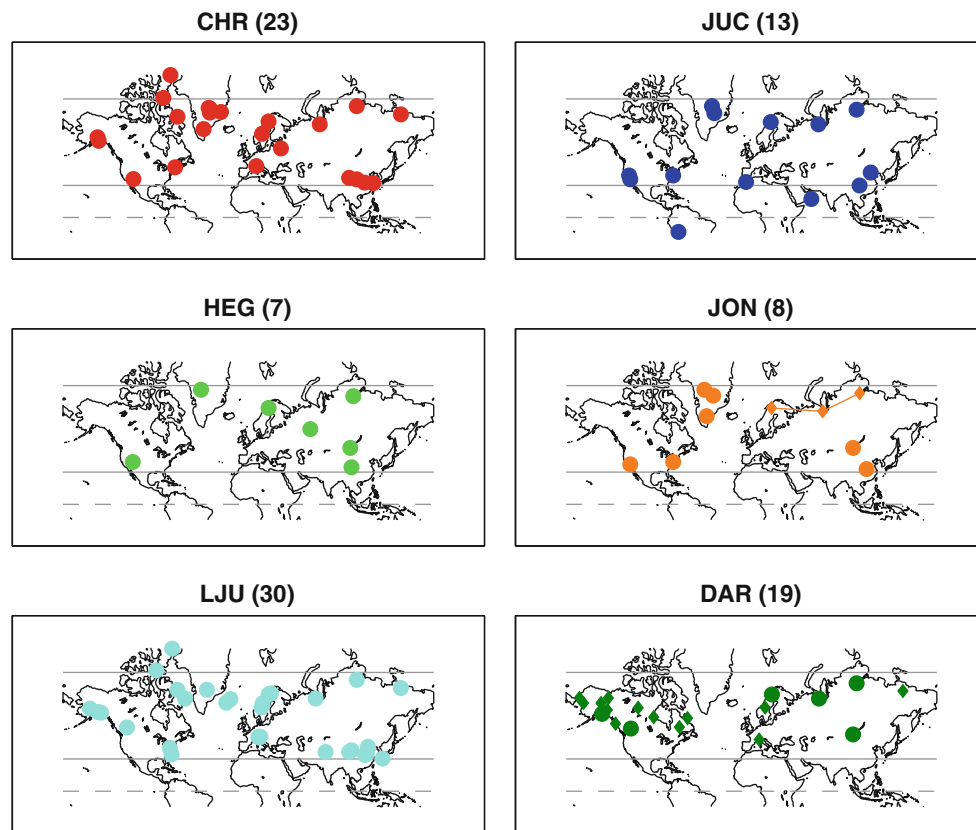


Fig. 1 Location of the local/regional temperature proxy series used in each of the six selected Northern Hemisphere-scale temperature reconstructions by Christiansen and Ljungqvist (2011) (CHR), Juckes et al. (2007) (JUC), Hegerl et al. (2007) (HEG), Jones and Mann (2004) (JON), Ljungqvist (2010) (LJU), D'Arrigo et al. (2006) (DAR). The numbers within parentheses indicate the number of regional/local proxy series used by each author, according to their own definitions. In the case of JON, the three sites marked with

diamonds and connected with *lines* were treated by them as a single regional composite. In the case of DAR, the 13 sites marked with *diamonds* denote proxy records that do not reach back to AD 1000, but start in different years between AD 1140 and 1686. In some cases where the size of the region that a proxy series represents is larger than the *symbol*, the approximate central location is marked. The latitudes 30°N and 75°N are indicated with *solid lines* and the equator with *dashes*

et al. 2007), all areas north of 30°N (Ljungqvist 2010; Christiansen and Ljungqvist 2011), or land-only areas north of 20°N (D'Arrigo et al. 2006)/30°N (Hegerl et al. 2007). There has also been a great deal of debate (Bürger et al. 2006; Christiansen et al. 2009; Ammann et al. 2010) as to how the final calibration against the instrumental target should be done, a discussion that also involves the 'climate field reconstruction' methods (Mann et al. 2009), where the full temperature field is first reconstructed and then a large-scale average is computed.

The range of reconstructions used here consists of Jones and Mann (2004); D'Arrigo et al. (2006); Juckes et al. (2007); Hegerl et al. (2007); Ljungqvist (2010); Christiansen and Ljungqvist (2011), henceforth labelled JON, DAR, JUC, HEG, LJU, CHR respectively. Five of these six reconstructions are multi-proxy compilations that include tree-ring data, whereas DAR is based solely on tree-ring data. The number and locations of the individual local or regional proxy series used by the different authors is illustrated in

Fig. 1. There it can be seen that nearly all the proxy data used are located on land between 30°N and 75°N. Thus, although none of the reconstructions were originally calibrated against land-only temperatures averaged between these latitudes, we judge that this is a valid common target against which all reconstructions can be re-calibrated for the sake of consistency. Many individual proxy series used by the different authors, in particular the tree-ring records, largely reflect temperatures in the warm part of the year; which would motivate calibration against temperatures averaged over the summer or the summer half-year, such as in Briffa et al. (2001). However, none of the six reconstructions used here targeted summer mean temperatures in their original calibrated forms, but rather annual averages. Further details of the various data selections and reconstruction methods used by the original authors are given in our supplementary material.

Note that the equally important 'RegEM' reconstructions by Mann et al. (2008, 2009) could not be included in

the present analysis as they consist of instrumental data during the calibration period, rendering them inappropriate for the Sundberg et al. (2012) statistical framework which requires re-calibration of all reconstructions (see Sect. 2.3 below). This is not possible without access to a proxy-based reconstruction in parallel with an instrumental record.

2.2 Instrumental data

As a common calibration target, chosen to be suitable for all six reconstructions, we employed the land-only temperatures averaged over 30°–75°N derived from the CRUTEM3 dataset (Brohan et al. 2006). This record starts in 1850, but we only use data back to 1880 due to a larger uncertainty in earlier years. For the calibration, we use non-overlapping 10-year mean temperatures within the period 1880–1959, which is the longest feasible period to be consistently used with all six reconstructions (further motivation and discussion can be found in the supplementary text).

We consider both the 6-month warm season April–September (AMJJAS) and annual averages, though focus on the former here for reasons discussed in Sect. 2.3. The error variance of the instrumental temperature record, as allowed in the framework of Sundberg et al. (2012), was estimated by combining the uncertainties caused by the station and measurement errors, sampling within grid-boxes and the limited coverage due to grid-boxes with missing data [derived by Brohan et al. (2006)]. These errors together are found to represent roughly 10 % of the total variance in observed 10-year average surface temperatures within the calibration period (see supplementary material for a detailed explanation of the calculations).

2.3 Proxy (re-)calibration ensemble

The Sundberg et al. (2012) statistical framework explicitly states how the proxy data should be calibrated in order to obtain an unbiased ranking of simulations. If the instrumental error variance is zero or negligible, then the calibration should be done by regressing the proxy on the instrumental data and then inverting the slope—a method known as ‘classical calibration’ in the statistical literature (Osborne 1991). If instrumental errors are known, and non-negligible, then we have an errors-in-variables situation (Fuller 1987; Cheng and van Ness 1999) and a correction of the regression slope has to be made. In both cases, the resulting calibrated series should have the correct variance of the true (but unobservable) temperature signal with the error (noise) variance superimposed. Thus, the resulting calibrated proxy has more variance than the instrumental data. Sundberg et al. (2012) provides a detailed discussion

of the theoretical arguments for why the calibration should be undertaken in this way. To our knowledge, this type of calibration has previously been applied only once with hemispheric-scale ‘index-type’ temperature reconstructions, namely by Hegerl et al. (2007), although the problem has been discussed more recently also, e.g. by Ammann et al. (2010), Christiansen (2011), Tingley et al. (2012).

To investigate the degree to which comparisons of the simulations with the Northern Hemispheric-scale reconstructions are affected by uncertainty due to the choice of calibration dataset, we created an ensemble with 100 calibration alternatives for each of the six reconstructions, by randomly sampling with replacement pairs of instrumental and proxy data from the calibration period. This approach accounts for calibration uncertainty in a similar way to the investigation of Frank et al. (2010), where several reconstructions are calibrated in a consistent way but with allowance for randomness in the choice of calibration dataset.

For all six reconstructions, the correlation coefficients between the proxy and the instrumental series were found to be strong (often larger than 0.9, see supplementary material). However, as the calibration data set was very small (with only eight pairs of 10-year mean temperatures and proxy values in the calibration period) it was naturally the case that some of the re-samplings led to very weak (occasionally negative) correlations between the instrumental series and the proxy. As insignificant, or even negative, correlations would contradict the original authors’ conclusions that their reconstructions are valid proxies for large-scale temperatures, we dismissed the few occasions when re-samplings led to correlations below 0.2. In the rare cases where this did occur, we generated a new random re-sampling to ensure that physically meaningless re-calibrations are avoided.

The generally very strong correlations, however, often turned out to be in conflict with the independently estimated instrumental error variance, given the assumptions about uncorrelated errors made in Sundberg et al. (2012). For example, if the instrumental noise variance accounts for 10 % of the total variance of the observed temperatures, then the squared correlation with the proxy cannot in theory be larger than 0.90 (again, see supplementary material for details). Whenever these conflicts occurred, we decided to reduce the estimated instrumental error variance and set it equal to the largest possible value that is not in conflict with the theory and assumptions made about the data. These conflicts were found to be sufficiently less of a problem for AMJJAS average temperatures, compared with annual averages, hence we decided to focus here on the AMJJAS season. Note that annually averaged results were calculated for all cases as well and can be viewed in the supplementary material.

2.4 COSMOS simulations

The Max Planck Institute developed the COSMOS Millennium Activity simulations (Jungclaus et al. 2010) using their Earth System Model (MPI-ESM), which comprise both atmospheric [ECHAM5—T31 (3.75)] and ocean (MPIOM) models, as well as those for land vegetation (JSBACH) and ocean biogeochemistry (HAMOCC). The ocean and atmospheric models are coupled daily without the use of flux correction. The COSMOS simulations comprise a 3,000-year long CTRL simulation (with orbital conditions set to 800 AD and pre-industrial greenhouse gas levels) as well as several simulations for the last millennium with either low or high amplitudes of estimated TSI forcing series driving their climates. Two full-forcing ensembles were generated (using different ocean initial conditions) using either the low (E1) or high (E2) solar forcing histories in combination with other known principal drivers of climate, namely aerosols (volcanic and non-volcanic), greenhouse gases (CO₂, N₂O, CH₄), orbital and land-use changes (Jungclaus et al. 2010). The low and high magnitude solar forcing series are the reconstructions of Krivova et al. (2007) and Bard et al. (2000) respectively, exhibiting decreases by 0.1 and 0.25 % when comparing the Maunder Minimum period to present values. The low solar being in agreement with the largest amplitudes used in PMIP3 and the high solar series representing a more commonly held value from the late 1990s. Note that the Krivova et al. (2007) and Bard et al. (2000) TSI reconstructions do not only differ in terms of their amplitudes; there are also some differences in the shapes of their temporal evolution, which will to some degree influence the results of our comparisons of simulated and reconstructed temperatures. Seeing as only one physical climate *model* (as opposed to ensemble member with different initial climate state) is used in this analysis, we may be under-representing the uncertainty of internal dynamics in the climate system. Additionally, the sensitivity of simulated temperature response to solar variations can be model-dependent. The COSMOS simulations estimated global temperature change per Wm⁻² (TSI) is 0.15 K/(Wm⁻²) in response to the 11-year solar cycle (Jungclaus et al. 2010), which lies within the range of recent estimates [0.1–0.2 K/(Wm⁻²); Camp and Tung (2007), Lean and Rind (2008)], if we consider this analysis to be applicable for other GCMs.

2.5 Statistical method

The statistical framework of Sundberg et al. (2012) involves two statistical tests: firstly, an initial correlation test to establish whether a given simulation is able to explain any of the temporal variation in the observed temperature data; secondly, a distance measure test created to quantify whether a given simulation is significantly closer, or not, to the ‘true’ temperatures for a given target region than unforced control simulations. As the true past

temperatures are not perfectly known, they have to be estimated by instrumental data or proxy data. The goal of this methodology, given that the correlation test statistics (U_R) are significantly positive (indicating that the simulations and the instrumental/proxy data share a common forced signal), is to rank competing climate model simulations in terms of how close they are to the un-observable true surface temperature variations using the distance-based test statistic U_T [see Sundberg et al. (2012); Sect. 2–8 for a complete derivation].

Both test statistics involve the calculation of a weighting at every time point in the comparison, the purpose of which is to give higher weight in periods where the observed/reconstructed temperature data have higher statistical precision (and thereby also smaller variance, as a result of how the calibration is made). As the instrumental data have higher precision than proxy data, the methodology states that instrumental data should be used whenever available and be replaced by the less precise proxy data only in the pre-instrumental period. Hence, in this analysis, the time series being used for comparison with the simulations consist of instrumental data after 1880 and the (re-)calibrated proxy series before 1880. Consequently, more weight is given to data in the post-1880 period (whenever this is used in the analysis), although the difference in weights is mostly small as the calibration generally revealed strong correlations between the proxy and instrumental data.

The null hypothesis of the U_T statistical test is that a (forced) simulation is equivalent to CTRL simulations. For this statistic (as well as for U_R) to have a given type I error level (e.g. 0.05), the unforced CTRL simulations should be represented by white noise. It is well known that the climate system has a great deal of autocorrelation, however on longer time scales it may be the case that the autocorrelation of internal variability behave insignificantly enough to be approximated by white noise. Here, the comparison was conducted on 20-year non-overlapping averages, as this time resolution was found to be the highest we could comfortably use before the COSMOS CTRL hemispheric-scale temperatures displayed significant autocorrelation [see our supplementary material and also discussions in Sundberg et al. (2012) and Hind et al. (2012)].

The framework of Sundberg et al. (2012) includes a choice concerning how to average data from ensemble simulations that share the same forcing and only differ by their initial conditions, such as each of the COSMOS E1 and E2 ensembles. Here, we choose the so-called ‘inside’ averaging described in Appendix A of Sundberg et al. (2012), as found by Hind et al. (2012) (their Appendix A) to be more effective than the alternative ‘outside’ averaging. This means that the U_T test statistics here involve the averaging of time sequences for the ensemble members so as to get a single representative fully-forced hemispheric ensemble-

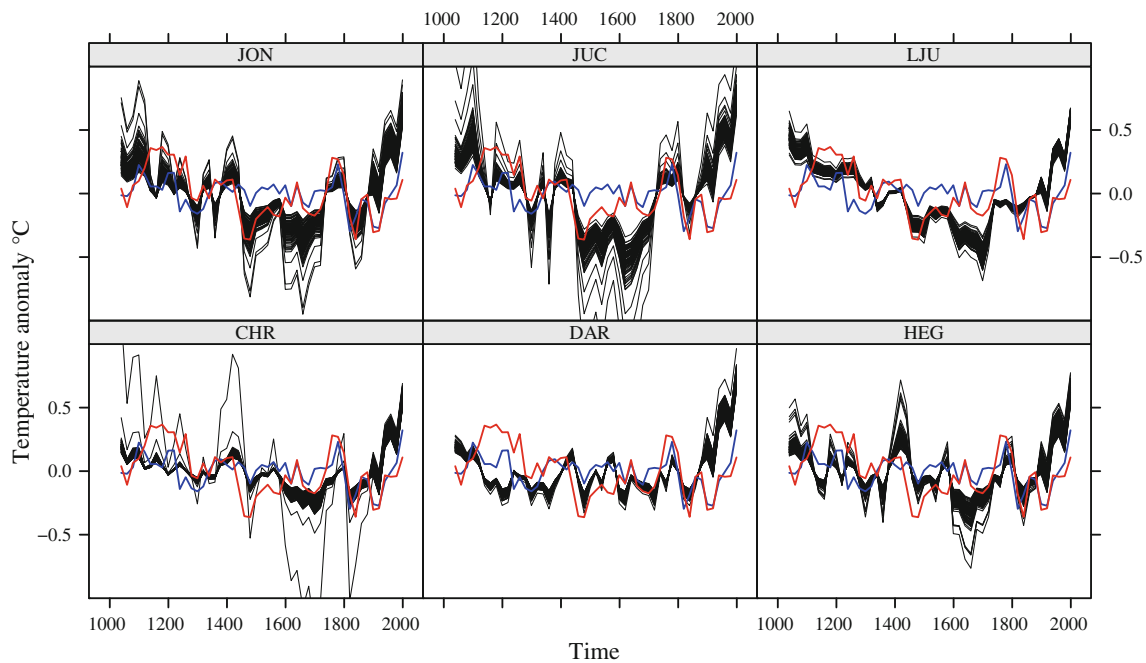


Fig. 2 Surface temperature anomalies ($^{\circ}\text{C}$) for six millennial hemispheric-scale reconstructions re-calibrated against average 30° – 75°N AMJJAS temperatures (*black lines*, one hundred calibration replicants each) compared with the corresponding simulated E1 (*blue*) and E2 (*red*) ensemble-mean temperatures from the COSMOS simulations by Jungclauss et al. (2010), identically repeated in *each panel*. The

mean simulation time series, for each of E1 and E2 respectively. The 3,000-year CTRL simulation was converted to a similar ensemble-mean series, by first splitting the whole simulation into three millennial segments, and then averaging over the three members to obtain one 1,000-year long CTRL series. Note that for a given analysis period (say 1000–1850 or 1400–2000), the mean temperature for the whole of this period is always subtracted separately from the instrumental/proxy data and the simulation data. Hence, the distance measures only consider deviations from the mean in the actual analysis period, therefore any systematic model biases will not affect the ranking of simulations. We should also say that the framework can be utilized for jointly comparing multiple individual proxy locations, however this analysis will only deal with hemispheric-scale averages.

3 Results

3.1 Simulated versus reconstructed average 30° – 75°N AMJJAS temperatures

Figure 2 shows the ensemble-average E1 (blue) and E2 (red) temperature series for the 1000–2000 period, with their long-term means removed, against the corresponding reconstructed 30° – 75°N temperature series (black). For each of the six original reconstructions, there is an

reconstructions are identified by their author abbreviations as in Fig. 1. All data are shown as 20-year non overlapping mean temperatures (connected by *lines*) with the mean over the entire period removed. Note: all reconstruction time series consist of identically the same instrumental temperature data after 1880, but of alternatively calibrated proxy data before this year

ensemble with one hundred alternatives designed to give some impression of calibration uncertainty (see Sect. 2.3). The high solar E2 simulations show larger multi-centennial variations than the low solar E1, leading to consistently higher warming during the medieval period in E2, with a peak around 1200 and relatively colder temperatures from 1400–1600. Note that all reconstruction time series consist of instrumental temperature data after 1880 and of the alternatively calibrated proxy data before this year. Note also that due to the calibration procedure, the proxy-data part of the records have an exaggerated variance as the data should consist of a correctly estimated variance of the hypothetical true temperature variations with the noise term superimposed. The weaker the correlations with the instrumental temperatures, the larger the variance in the corresponding calibration replicants. This may complicate any direct visual comparison with the simulation time series, but the effect is taken care of in the calculation of the two statistical measures (U_R , U_T) as smaller weights are given where the proxy data has a weaker correlation with temperature.

3.2 Distributions of U_R and U_T test statistics

Through the U_R and U_T test statistics, the Sundberg et al. (2012) framework allows us to compare the simulations with the reconstructions in a consistent and quantitative

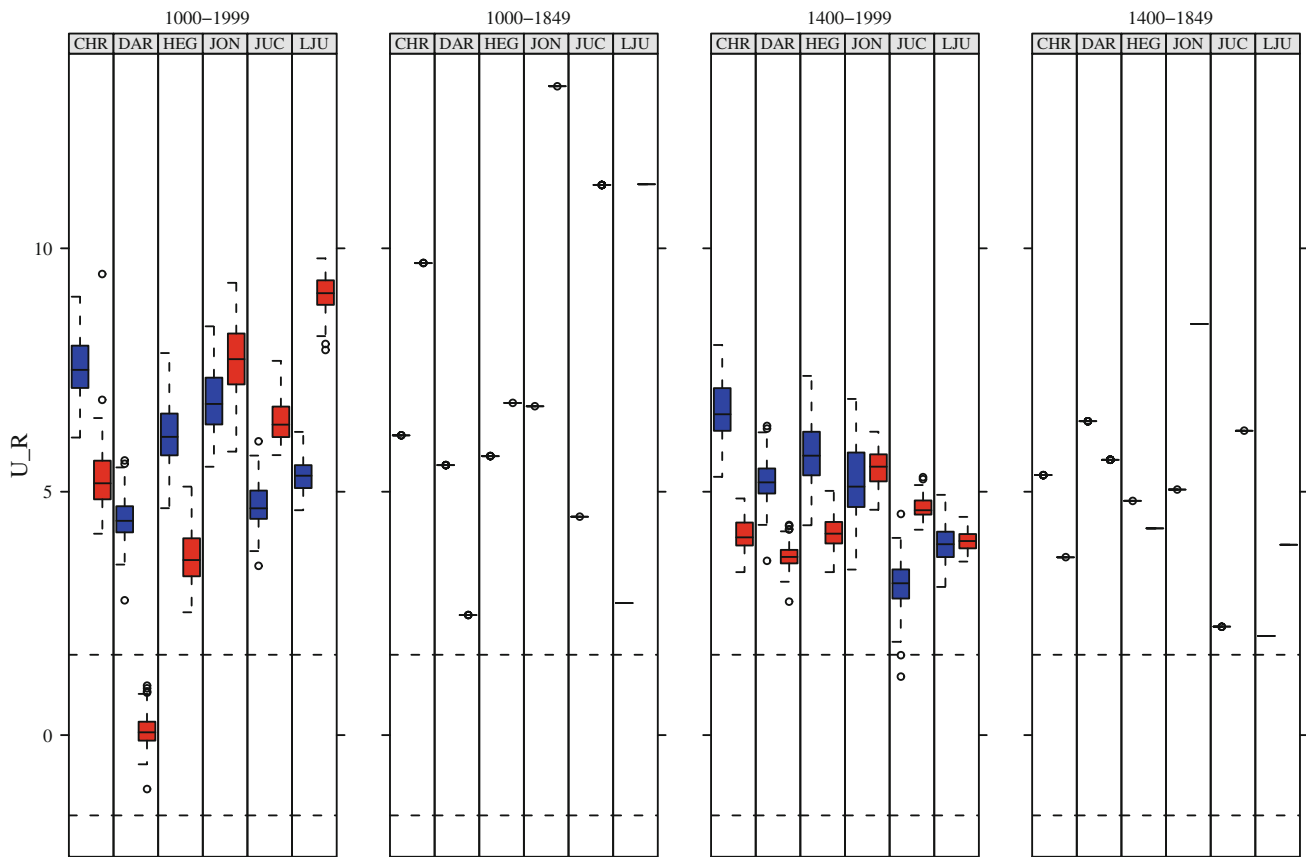


Fig. 3 The U_R correlation test statistics comparing the E1 (blue) and E2 (red) ensemble simulation mean time-series with the (re-)calibration ensembles of the six hemispheric reconstructions (i.e. their hundred replicants), over the period 1000–1999 (far left), 1000–1849 (left), 1400–1999 (right), 1400–1849 (far right). The capitalized reconstruction name labels are the same as those used in Fig. 1. The

dashed lines indicate statistical significance at the 0.05 two-sided levels. The filled boxes represent the lower quartile (bottom edge), median value (horizontal black line) and upper quartile (upper edge). The whiskers represent the lowest and highest data points within 1.5 interquartile range of the lower and upper quartiles, respectively. The circles represent outlying values

way. Figure 3 represents the distribution of U_R test statistics for both the E1 (blue) and E2 (red) ensemble simulations when they are compared with all members of each proxy (re-)calibration ensemble, illustrated with box-plots. Each column within the panels represents the different reconstructions; CHR, DAR, HEG, JON, JUC, LJU from left to right. The four panels represent four different analysis periods, which were conducted in order to see if the results are sensitive to the choice of period made. These consist of 1000–1999, 1000–1849, 1400–1999, 1400–1849; motivated by previous observations of a notable mismatch between simulated and reconstructed temperatures approximately before AD 1400 (Servonnat et al. 2010; González-Rouco et al. 2011). We also wish to allow the exclusion or inclusion of the post-1850 period, it being strongly affected by anthropogenic factors, including the uncertain effect of aerosols which may complicate model versus data comparisons (Schwartz et al. 2007).

An inspection of Fig. 3 for these four periods indicates that both E1 and E2 simulations have significant positive

correlations with all reconstructions (with the exception of E2 against DAR in 1000–1999). Note that when data after 1850 is included in the comparison there is a spread in U_R values but not when this period is excluded. This is because when the reconstructions consist of merged proxy data and instrumental data, the individual members of each (re-) calibration ensemble are truly different time series, whereas when only proxy data are used the ensemble members differ only by a multiplicative factor; i.e. they are simply scaled versions of one and the same time series. Therefore, in the latter situation, their correlations with the simulations are always the same, regardless of how they were calibrated.

The general implication of Fig. 3 is that we observe a significant positive correlation between both the low and high solar full-forcing ensemble simulations and reconstructed surface temperature. These results do not of themselves indicate that both low and high solar forcing series match the target data in the fully-forced simulations, as there is no way of knowing whether the significant

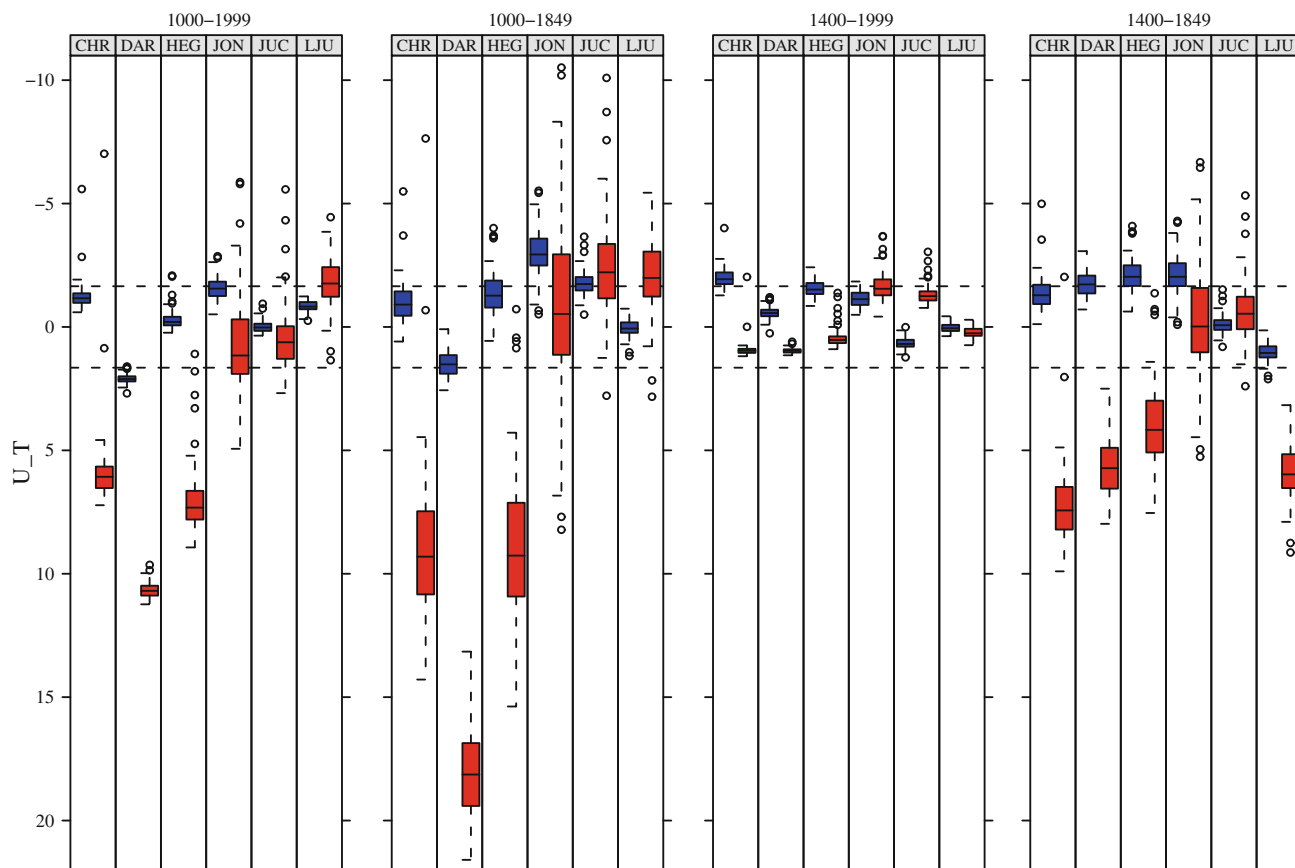


Fig. 4 As for Fig. 3, but for the U_T test statistic. Note that the y-axes are reversed, as a more negative value (*upwards* in the *graphs*) indicates a better performance (i.e. a simulation being closer to the target series)

correlations are due to the solar forcing alone, or if they are caused by one of the other forcings or by the combination of several forcings. The important result, however, is that we can conclude with statistical certainty that the simulations and the reconstructions share a common forced signal. This in turn means that a ranking of the E1 and E2 simulation ensembles using the U_T statistic is meaningful.

Figure 4 represents the U_T distributions for the E1 and E2 simulations and has the same panel format as Fig. 3. Note that a negative U_T value indicates a given simulation to be closer to the target series than the CTRL simulations are, and significantly closer (at the two-sided 0.05 level) if lying above the upper horizontal dashed line. A simulation is significantly more distant than CTRL simulations if a test value lies below the lower dashed line. Neither the E1 or E2 simulations can be said to consistently outperform the controls during any of the periods. Neither is it possible to consistently rank E1 and E2 against each other considering all reconstructions and time periods. Nevertheless, the E1 simulations perform better than E2 in 15 of the 24 cases, in so far as their inter-quartile ranges do not overlap and E1 have the more negative or less positive (i.e. ‘better’) test

values. Viewed in a similar way, E2 outperforms E1 in only 3 of the 24 cases.

What is clear from these results, however, is that the E2 ensemble averages are sometimes significantly further away from the target data than control simulations in all periods with the exception of 1400–1999, where the medieval period is excluded from the analysis. Though it should be noted that this is not consistently the case for all the Northern Hemispheric-scale reconstructions, it is the case for CHR, DAR and HEG. Regarding E1, it is only when these simulations are compared with DAR, and only when the pre-1400 data are included, that the results point to a performance possibly worse than for CTRL simulations—whilst E2 performs considerably worse in this case. Thus, in terms of ranking the two simulation ensembles, it certainly seems as though the E1 simulations are far better than the E2 simulations comparing the CHR, DAR and HEG reconstructions, whilst the differences are not so clear in the other reconstructions. In summary, none of the six reconstructions consistently support E2 being closer than E1 to the real past hemispheric-scale AMJJAS temperatures, whilst three of the reconstructions consistently support E1 over E2.

4 Discussion and conclusions

Using the statistical framework of Sundberg et al. (2012), we have compared Northern Hemispheric-scale temperatures in a set of millennial-length GCM simulations (Jungclaus et al. 2010) driven by two solar forcing histories with different magnitudes, but the same orbital, volcanic and anthropogenic climate forcing, with six well-known temperature reconstructions from proxy data for the last millennium. Three of these reconstructions (Jones and Mann 2004; D'Arrigo et al. 2006; Hegerl et al. 2007) were used in the latest IPCC report (Jansen et al. 2007), while the other three (Juckes et al. 2007; Ljungqvist 2010; Christiansen and Ljungqvist 2011) were published later. Together, the six well represent a variety of choices among authors to select and combine proxy data into temperature reconstructions. To use the reconstructions in a consistent way, we re-calibrated all of them against April–September mean temperatures averaged over land areas between 30°–75°N. Whilst we cannot conclusively say which of the two solar forcing alternatives—the “low”-amplitude series of Krivova et al. (2007) or the “high”-amplitude series by Bard et al. (2000)—provides the best match with the reconstructed temperatures, there is clear evidence suggesting that the latter case does not match the reconstructions well. Thus, we find stronger statistical support for the lower Krivova et al. (2007) solar forcing amplitude; in accordance with the current view taken within the PMIP3 consortium (Schmidt et al. 2011), namely that TSI values during the Maunder Minimum period (AD 1645–1715) were between 0.04 and 0.1 % smaller compared to the present (including the temporally extensive TSI reconstruction of Steinhilber et al. (2009) having a corresponding reduction of 0.07 %). Contrastingly, we find no consistent support for the alternative Bard et al. (2000) solar forcing series, which has a 0.25 % TSI reduction in the Maunder Minimum, although one or two of the six reconstructions favour this alternative in some sub-periods investigated.

Although this research has focused on April–September mean temperatures, we also conducted the analysis on annually averaged temperatures, where it was found that the ranking of the simulations with high and low solar forcing is essentially unchanged. However, a generally much better performance of both simulation types against all six reconstructions is seen if annual-mean temperatures are used when data before AD 1400 is excluded from the analysis (see supplementary Figs. 10 and 11). We have not investigated the reasons for this, but we should point out that all reconstructions were originally calibrated against annual mean temperatures by their authors.

As we have only statistically compared simulations made with a single climate model driven by two alternative

solar forcings with a particular selection of temperature reconstructions, we are not in a position to quantify how large the multi-centennial TSI variations have been in the past millennium. Such a judgement can not strictly be made by comparing forced simulations with climate reconstructions, due to the simultaneous uncertainties in the solar forcing (Gray et al. 2010), climate sensitivity (Knutti and Hegerl 2008) and past climate variations (Jansen et al. 2007). Nor have we addressed potential complexities in differentiating between dynamical and thermal atmospheric process responses to solar forcing (Lean and Rind 2008), such as the east North Atlantic winter circulation pattern linked by Woolings et al. (2010) to the the open solar flux measure of solar activity.

We can only judge here that, given our selection of proxy-based temperature reconstructions, it is more likely that the lower-amplitude solar forcing reconstruction of Krivova et al. (2007) is realistic compared with the larger-amplitude solar reconstruction of Bard et al. (2000), although it must be remembered that such a result must be dependent on the sensitivity of climate model used. It is evident that the results of our model versus data comparison are strongly affected by the choice of individual proxy data and how they were combined into hemispheric-scale reconstructions by the original investigators. The range of proxy selections in these reconstructions may affect the aforementioned regional response to solar forcing, which suggests a dedicated regional-scale analysis could provide more informative results to those presented here.

Thus, although this problem was pointed out in the latest IPCC report (Jansen et al. 2007), greater consensus is still needed on climate history over the last millennium, particularly if modelling can be utilized to test ideas regarding the forced and unforced variability of the climate system.

Acknowledgments This research was funded by the Swedish Research Council (90751501 and B0334901). We thank Philip Brohan for providing data and help with how to estimate the error variance in CRUTEM3 data. We also thank our Stockholm colleagues Rolf Sundberg and Gudrun Brattström for their statistical expert guidance, as well as Johan Jungclaus of the Max Planck Institute for providing the COSMOS data.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

- Ammann CM, Genton MG, Li B (2010) Technical note: correcting for signal attenuation from noisy proxy data in climate reconstructions. *Clim Past* 6:273–279

- Bard E, Raisbeck G, Yiou F, Jouzel J (2000) Solar irradiance during the last 1200 years based on cosmogenic nuclides. *Tellus B* 52:985–992
- Braconnot P, Harrison SP, Kageyama M, Bartlein PJ, Masson-Delmotte V, Abe-Ouchi A, Otto-Bliesner B, Zhao Y (2012) Evaluation of climate models using palaeoclimatic data. *Nat Clim Chang* 2:417–424. doi: [10.1038/nclimate1456](https://doi.org/10.1038/nclimate1456)
- Briffa KR, Osborn TJ, Schweingruber FH, Harris IC, Jones PD, Shiyatov SG, Vaganov EA (2001) Low-frequency temperature variations from a northern tree ring density network. *J Geophys Res* 106:2929–2941
- Brohan P, Kennedy JJ, Harris I, Tett SFB, Jones PD (2006) Uncertainty estimates in regional and global observed temperature changes: a new data set from 1850. *J Geophys Res* 111. doi: [10.1029/2005JD006548](https://doi.org/10.1029/2005JD006548)
- Bürger G, Fast I, Cubasch U (2006) Climate reconstruction by regression—32 variations on a theme. *Tellus* 58A:227–235. doi: [10.1111/j.1600-0870.2006.00164.x](https://doi.org/10.1111/j.1600-0870.2006.00164.x)
- Camp CD, Tung KK (2007) Surface warming by the solar cycle as revealed by the composite mean difference projection. *Geophys Res Lett* 34:L14703
- Cheng CL, van Ness JW (1999) *Statistical regression with measurement error*. Arnold Publishers, London
- Christiansen B (2011) Reconstructing the NH mean temperature: can underestimation of trends and variability be avoided? *J Clim* 24:674–692. doi: [10.1175/2010JCLI3646.1](https://doi.org/10.1175/2010JCLI3646.1)
- Christiansen B, Ljungqvist FC (2011) Reconstruction of the extra-tropical NH mean temperature over the last millennium with a method that preserves low-frequency variability. *J Clim* 24:6013–6034. doi: [10.1175/2011JCLI4145.1](https://doi.org/10.1175/2011JCLI4145.1)
- Christiansen B, Schmith T, Thejll P (2009) A surrogate ensemble study of climate reconstruction methods: stochasticity and robustness. *J Clim* 22:951–976. doi: [10.1175/2008JCLI2301.1](https://doi.org/10.1175/2008JCLI2301.1)
- D'Arrigo RD, Wilson R, Jacoby G (2006) On the long-term context for late twentieth century warming. *J Geophys Res* 111. doi: [10.1029/2005JD006352](https://doi.org/10.1029/2005JD006352)
- Eddy JA (1976) The maunder minimum. *Science* 192:1189–1202
- Feulner G (2011) Are the most recent estimates for Maunder Minimum solar irradiance in agreement with temperature reconstructions?. *Geophys Res Lett* 38 doi: [10.1029/2011GL048529](https://doi.org/10.1029/2011GL048529)
- Frank DC, Esper J, Raible CC, Büntgen U, Trouet V, Stocker B, Joos F (2010) Ensemble reconstruction constraints on the global carbon cycle sensitivity. *Nature* 463:527–530. doi: [10.1038/nature08769](https://doi.org/10.1038/nature08769)
- Fuller WA (1987) *Measurement error models*. Wiley, New Jersey
- González-Rouco FJ, Fernández-Donado L, Raible CC, Barriopedro D, Luterbacher J, Jungclauss JH, Swingedouw D, Servonnat J, Zorita E, Wagner S, Ammann CM (2011) Medieval climate anomaly to little ice age transition as simulated by current climate models. In: Xoplaki E, Fleitmann D, Diaz H, Gunten L, Kiefer T (eds) *PAGES news: medieval climate anomaly*. Läderach AG, Bern, Switzerland
- Gray LJ, Beer J, Geller M, Haigh JD, Lockwood M, Matthes K, Cubasch U, Fleitmann D, Harrison G, Hood L, Luterbacher J, Meehl GA, Shindell D, van Geel B, White W (2010) Solar influence on climate. *Rev Geophys* 48:1–53. doi: [10.1029/2009RG000282](https://doi.org/10.1029/2009RG000282)
- Hegerl GC, Crowley TJ, Hyde WT, Frame DJ (2006) Climate sensitivity constrained by temperature reconstructions over the past seven centuries. *Nature* 440:1029–1032. doi: [10.1038/nature04679](https://doi.org/10.1038/nature04679)
- Hegerl GC, Crowley TJ, Allen M, Hyde WTNPH, Smerdon J, Zorita E (2007) Detection of human influence on a new, validated 1500-year temperature reconstruction. *J Clim* 20:650–666. doi: [10.1175/JCLI4011.1](https://doi.org/10.1175/JCLI4011.1)
- Hind A, Moberg A, Sundberg R (2012) 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](https://doi.org/10.5194/cp-8-1355-2012)
- Jansen E, Overpeck J, Briffa KR, Duplessy JC, Joos F, Masson-Delmotte V, Olago D, Otto-Bliesner B, Peltier WR, Rahmstorf S, Ramesh R, Raynaud D, Rind D, Solomina O, Villalba R, Zhang D (2007) Paleoclimate. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) *Climate change 2007: the physical science basis*. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA
- Jones PD, Mann ME (2004) Climate over past millennia. *Rev Geophys* 42:RG2002. doi: [10.1029/2003RG000143](https://doi.org/10.1029/2003RG000143)
- Jones PD, Briffa KR, Osborn TJ, Lough JM, van Ommen TD, Vinther BM, Luterbacher J, Wahl ER, Zwiers FW, Mann ME, Schmidt GA, Ammann CM, Buckley BM, Cobb KM, Esper J, Goosse H, Graham N, Jansen E, Kiefer T, Kull C, Küttel M, Mosley-Thompson E, Overpeck JT, Riedwyl N, Schulz M, Tudhope AW, Villalba R, Wanner H, Wolff E, Xoplaki E (2009) High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects. *Holocene* 19:3–49. doi: [10.1177/0959683608098952](https://doi.org/10.1177/0959683608098952)
- Juckes MN, Allen MR, Briffa KR, Esper J, Hegerl GC, Moberg A, Osborn TJ, Weber SL (2007) Millennial temperature reconstruction intercomparison and evaluation. *Clim Past* 3:591–609
- Jungclauss JH, Lorenz SJ, Timmreck C, Reick CH, Brovkin V, Six K, Segschneider J, Giorgetta MA, Crowley TJ, Pongratz J, Krivova NA, Vieira LE, Solanki SK, Klocke D, Botzet M, Esch M, Gayler V, Haak H, Raddatz TJ, Roeckner E, Schnur R, Widmann H, Claussen M, Stevens B, Marotzke J (2010) Climate and carbon-cycle variability over the last millennium. *Clim Past* 6:723–737. doi: [10.5194/cp-6-723-2010](https://doi.org/10.5194/cp-6-723-2010)
- Knutti R, Hegerl GC (2008) The equilibrium sensitivity of the earth's temperature to radiation changes. *Nat Geosci* 1:735–743
- Krivova NA, Balmaceda L, Solanki SK (2007) Reconstruction of solar total irradiance since 1700 from the surface magnetic flux. *Astron Astrophys* 467:335–346. doi: [10.1051/0004-6361:20066725](https://doi.org/10.1051/0004-6361:20066725)
- Lean JL, Rind DH (2008) How natural and anthropogenic influences alter global and regional surface temperatures: 1889 to 2006. *Geophys Res Lett* 35:L18 701. doi: [10.1029/2008GL034864](https://doi.org/10.1029/2008GL034864)
- Ljungqvist FC (2010) A new reconstruction of temperature variability in the extra-tropical Northern Hemisphere during the last two millennia. *Geografiska Annaler* 92A:339–351
- Lockwood M (2011) Shining a light on solar impacts. *Nat Clim Chang* 1:98–99
- Lockwood M (2012) Solar Influence on global and regional climates. *Survey Geophys* 33:503–534
- Mann ME, Zhang Z, Hughes MK, Bradley RS, Miller SK, Rutherford S, Ni F (2008) Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. *PNAS* 105:13 252–13 257
- Mann ME, Zhang Z, Rutherford S, Bradley RS, Hughes MK, Shindell D, Ammann C, Faluvegi G, Fensholt N (2009) Global signatures and dynamical origins of the little ice age and medieval climate anomaly. *Science* 326:1256–1260
- Osborne C (1991) Statistical calibration: a review. *Int Stat Rev/Revue Internationale de Statistique* 59:309–336
- Schmidt GA (2010) Enhancing the relevance of paleoclimate model/data comparisons for assessments of future climate change. *J Quat Sci* 25:79–87. doi: [10.1002/jqs.1314](https://doi.org/10.1002/jqs.1314)
- Schmidt GA, Junclaus JH, Ammann CM, Bard E, Braconnot PCTJDG, Joos F, Krivova NA, Muscheler R, Otto-Bliesner

- BL, Pongratz J, Shindell DT, Solanki SK, Steinhilber F (2011) 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](https://doi.org/10.5194/gmd-4-33-2011)
- Schrijver CJ, Livingston WC, Woods TN, Mewaldt RA (2011) The minimal solar activity in 2008–2009 and its implications for long-term climate modeling. *Geophys Res Lett* 38:L06 701
- Schwartz SE, Charlson RJ, Rodhe H (2007) Quantifying climate change—too rosy a picture? *Nat Rep Clim Chang* 1:23–24. doi:[10.1038/climate.2007.22](https://doi.org/10.1038/climate.2007.22)
- Servonnat J, Yiou P, Khodri M, Swingedouw D, Denvil S (2010) Influence of solar variability, CO₂ and orbital forcing between 1000 and 1850 AD in the IPSLCM4 model. *Clim Past* 6:445–460. doi:[10.5194/cp-6-445-2010](https://doi.org/10.5194/cp-6-445-2010)
- Shapiro AI, Schmutz W, Rozanov E, Schoell M, Haberreiter M, Shapiro AV, Nyeki S (2011) A new approach to the long-term reconstruction of the solar irradiance leads to large historical solar forcing. *Astron Astrophys* 529(A67):1–8. doi:[10.1051/0004-6361/201016173](https://doi.org/10.1051/0004-6361/201016173)
- Steinhilber F, Beer J, Froehlich C (2009) Total solar irradiance during the Holocene. *Geophys Res Lett* 36. doi:[10.1029/2009GL040142](https://doi.org/10.1029/2009GL040142)
- Sundberg R, Moberg A, Hind A (2012) 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](https://doi.org/10.5194/cp-8-1339-2012)
- Tingley MP, Craigmile PF, Haran M, Li B, Mannshardt E, Rajaratnam B (2012) Piecing together the past: statistical insights into paleoclimatic reconstructions. *Quat Sci Rev* 35:1–22. doi:[10.1016/j.quascirev.2012.01.012](https://doi.org/10.1016/j.quascirev.2012.01.012)
- Wahl ER, Smerdon JE (2012) Comparative performance of paleoclimate field and index reconstructions derived from climate proxies and noise-only predictors. *Geophys Res Lett* 39. doi:[10.1029/2012GL051086](https://doi.org/10.1029/2012GL051086)
- Woolings T, Lockwood M, Masato G, Bell C, Gray L (2010) Enhanced signature of solar variability in Eurasian winter climate. *Geophys Res Lett* 37:201–213
- Yoshimori M, Stocker TF, Raible CC, Renold M (2005) Externally forced and internal variability in ensemble climate simulations of the Maunder Minimum. *J Clim* 18:4253–4270