Clim. Past Discuss., 10, 2627–2683, 2014
www.clim-past-discuss.net/10/2627/2014/
doi:10.5194/cpd-10-2627-2014

Climate of the Past
Discussions
Open Access

Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium – Part 3: Practical considerations, relaxed assumptions, and using tree-ring data to address the amplitude of solar forcing

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

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

Received: 13 May 2014 – Accepted: 2 June 2014 – Published: 26 June 2014

Correspondence to: A. Moberg (anders.moberg@natgeo.su.se)

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

Abstract

Practical issues arise when applying a statistical framework for unbiased ranking of alternative forced climate model simulations by comparison with climate observations from instrumental and proxy data (Part 1 in this series). Given a set of model and observational data, several decisions need to be made; e.g. concerning the region that each proxy series represents, the weighting of different regions, and the time resolution to use in the analysis. Objective selection criteria cannot be made here, but we argue to study how sensitive the results are to the choices made. The framework is improved by the relaxation of two assumptions; to allow autocorrelation in the statistical model for simulated climate variability, and to enable direct comparison of alternative simulations to test if any of them fit the observations significantly better. The extended framework is applied to a set of simulations driven with forcings for the pre-industrial period 1000–1849 CE and fifteen tree-ring based temperature proxy series. Simulations run with only one external forcing (land-use, volcanic, small-amplitude solar, or large-amplitude solar), do not significantly capture the variability in the tree-ring data – although the simulation with volcanic forcing does so for some experiment settings. When all forcings are combined (using *either* the small- or large-amplitude solar forcing) including also orbital, greenhouse-gas and non-volcanic aerosol forcing, and additionally used to produce small simulation ensembles starting from slightly different initial ocean conditions, the resulting simulations are highly capable of capturing some observed variability. Nevertheless, for some choices in the experiment design, they are not significantly closer to the observations than when unforced simulations are used, due to highly variable results between regions. It is also not possible to tell whether the small-amplitude or large-amplitude solar forcing causes the multiple-forcing simulations to be closer to the reconstructed temperature variability. This suggests that proxy data from more regions and proxy types, or representing larger regions and other seasons, are needed for more conclusive results from model-data comparisons in the last millennium.

1 Introduction

While much of our knowledge about climate changes in the past emerge from evidence in various natural archives (Wanner et al., 2008; Jones et al., 2009), experiments with climate models help to understand physical mechanisms behind the observed changes and may also help constrain projections of future climate changes (Schmidt, 2010). The last millennium – prior to the onset of the industrial era around 1850 CE – provides an opportunity to test hypotheses about the role of external drivers, in particular orbital forcing, solar variability, volcanic aerosols, land use/land cover changes and variations in greenhouse gas levels, under climate conditions close to those of today (Jungclaus et al., 2010; Schmidt et al., 2011; Landrum et al., 2012; Fernández-Donado et al., 2013; Sueyoshi et al., 2013). A constantly growing number of proxy-based reconstructions and model-based simulations of past climate variations implies an increasing need for statistical methods for comparing data of the two kinds. Examples of this are found in data assimilation (Goosse et al., 2012; Widmann et al., 2010), detection and attribution studies (Hegerl et al., 2007, 2011) and estimation of climate sensitivity (Hegerl et al., 2006). So far, the available methods can however not account for the full complexity of the situation. There is thus a need for more theoretical work in this context.

Based on theoretical considerations and some assumptions, Sundberg et al. (2012, henceforth SUN12) formulated a statistical framework for evaluation of climate model simulations, primarily for the last millennium. Their goal was to develop tools for an unbiased ranking of a set of alternative forced simulations in terms of their hypothetical distance to the unobservable true temperature history, while using noisy proxy records and instrumental observations as approximations to the true temperature variability. In a companion pseudoproxy experiment, Hind et al. (2012) investigated the possibility of determining whether climate model simulations, driven by various external forcings, were able to explain past temperature variability in a situation when the "true" past temperature history, the forcing history and the proxy noise were known by design.

Here, we contribute further to the SUN12 work by discussing practical considerations arising when using real proxy data series that represent different seasons and regions of different size, having different lengths and statistical precision. To this end, we select a set of 15 tree-ring based temperature reconstructions, spread across North America, Eurasia and Oceania, which we use together with the same set of global climate model simulations (Jungclaus et al., 2010) as used by Hind et al. (2012). Another goal is to present an extension of the SUN12 framework, by relaxing two of its assumptions. This makes it possible, first, to allow some autocorrelation structures in the simulated temperatures and, second, to compare two alternative forced simulations directly to test if one of them matches the observed climate variations significantly better than the other. SUN12 assumed no autocorrelation and compared forced simulations only indirectly, by testing whether each of them matched the observed climate variations better than a reference simulation with constant forcing. Although full details of the SUN12 framework are already provided in their original work, we summarize essential aspects here for the benefit of the reader. The extended framework is explained in detail in two appendices. Much of our discussion deals with practical issues when applying the framework, for example concerning how to define geographical regions for model-data comparison, how to combine information representing different regions and seasons and how to decide upon the time resolution to use in the analysis.

This work also serves as a companion study to the hemispheric-scale analysis by Hind and Moberg (2013), who attempted to determine which of two alternative solar forcing histories that, in the presence of other forcings, provided the best fit between simulated (Jungclaus et al., 2010) and reconstructed temperatures. The two solar forcing histories had either a 0.1 % or 0.25 % change in Total Solar Irradiance since the Maunder Minimum period (i.e. 1645–1715 CE; c.f. Jungclaus et al., 2010; Lockwood, 2011; Schmidt et al., 2012; Fernández-Donado et al., 2013; Masson-Delmotte et al., 2013). As temperature proxies, Hind and Moberg (2013) used six hemispheric-scale temperature reconstructions; five based on multi-proxy compilations and one based solely on tree-ring data. They found, in most cases, a better match when the small-amplitude solar forcing was used, but results were not conclusive. This provokes questions regarding whether statistical model-data comparisons can tell which of the two alternative solar forcing histories is most correct. Tree-ring-based proxy data, which form the backbone of our knowledge of past temperature variations in the last millennium (Jones et al., 2009), have a potential to shed further light on this question. We apply the extended framework to the selected tree-ring data, in an attempt to examine whether more conclusive results can be obtained. The current article is, however, mainly intended as a methodology study where the model-vs-data analysis serves as a relevant demonstration case.

2 Statistical framework

To obtain a statistical methodology for ranking a set of plausible alternative forced simulations, SUN12 proposed a type of regional (or local) statistical model relating climate model simulation time series (x) via the (unobservable) true temperature sequence (τ) to instrumental temperature measurements and temperature proxy data series. The instrumental measurements and proxy data (used only when the former are missing) are here jointly called "observations" and denoted z.

Section 4 of SUN12 demonstrated that for unbiased ranking, the calibration of proxy data should aim at keeping the right size of the true temperature (τ) component in the proxy, with the noise component superimposed. Perfectly calibrated temperature proxy data (or instrumental data) could thus be written $z = \tau + \epsilon$, where ϵ is a measurement error type term (noise), uncorrelated with τ. The methodology allows the variance of the noise term ϵ to vary with time, depending on how the precision of the proxy varies. Thus, an entire z-sequence may be composed of different segments, each with its characteristic noise variance. Typically, one or more of the segments will consist of instrumental measurements, with noise variance generally expected to be smaller than in proxy segments.

Based on their statistical models, SUN12 developed two test statistics for comparing climate model simulation data with combined instrumental and proxy temperature data. First, before any attempt is made to rank alternative model simulations, it should be tested whether a statistically significant positive correlation can be seen between a simulation series and the observations, because otherwise there is no evidence that the simulations and the true temperature share any effect of the forcing under study.

For this correlation pre-test, SUN12 proposed (in their Sect. 8) a test statistic $R(x,z)$, based on a weighted regression of x on z. Note, however, that high correlation does not mean that the forcing (or the response to it) is of the right size in the climate model. In particular, a magnified forcing effect in x necessarily increases the correlation, also if the effect is exaggerated.

Assuming next that a correlation has been established, the distance between a simulation sequence and an observation sequence is formed, as a weighted mean squared distance, D_w^2:

$$D_\mathrm{w}^2(x,z) = \frac{1}{n}\sum_{i=1}^{n} w_i (x_i - z_i)^2.$$

Here, n is the number of time steps in the sequence. Section 5 in SUN12 explains how the weights w_i are calculated in practice. It is through the weights w_i that the framework allows a temporally varying statistical precision of the proxies. A time segment with low precision will have a small weight w_i.

When an ensemble of simulations driven by the same forcing (but differing in their initial conditions) is available, they should all be used in an averaging process. This can be done in two different ways. Either a D_w^2 value is computed for each simulation and the average of these values is used, or alternatively the averaging is made of the simulation time sequences in the ensemble, before a D_w^2 value is computed for this ensemble-mean time sequence. In SUN12, this was referred to as averaging “outside” and “inside”, respectively, and was discussed primarily in their Sect. 6 and Appendix A. The theoretical discussion showed that the latter should be more precise but with a risk for bias (which can, and must, be corrected for). The pseudoproxy study by Hind et al. (2012) confirmed that inside averaging can be more efficient than outside averaging in practice.

For comparison of different forced models, SUN12 used a normalized version of D_w^2, rather than D_w^2 itself. First, all D_w^2 were replaced by their differences from the D_w^2 of an unforced reference model (data x^*),

$$T(x, x^*, z) = D_\mathrm{w}^2(x, z) - D_\mathrm{w}^2(x^*, z). \tag{1}$$

Thus, a (relatively large) negative value of $T(x, x^*, z)$ is needed to show that a forced model fits the observations better than the unforced reference. The question "how large?" is answered by scale-normalizing the $T(x, x^*, z)$-value by its standard error (square root of variance), calculated under the null hypothesis

H_0: *The forced climate model is equivalent to the unforced reference model.*

The resulting test statistic is the measure proposed by SUN12 for ranking. In their Sect. 6, a formula is derived for the standard error of $T(x, x^*, z)$, depending only on the reference model output.

The test statistics for correlation and distance were derived under specific assumptions on the climate model simulations (whereas the true climate was arbitrary). For ranking, this need not be considered a problem, but regarded as test statistics we want them robust against model imperfections. In particular, we want to relax the following assumptions from SUN12:

- assumed lack of autocorrelation in the reference model simulations, i.e. these are statistically represented by white noise;
- truly unforced reference model, so in particular no joint time-varying forcing in x and x^*.

Concerning the first assumption, it is well-known that internal temperature variation can show autocorrelation, because the climate system acts as an integrator of the short-term weather variations (Hasselmann, 1976). Depending on time-scale (e.g. annual or decadal), short or long memory dominates. Vyushin et al. (2012) found that a first-order autoregressive representation (AR(1)) and a power law can be seen as lower and upper bounds for characterizing this persistence. Hind et al. (2012) and Hind and Moberg (2013) attempted to avoid this problem by using quite long time units in their studies (30 years and 20 years). This was empirically justified, as unforced temperatures in simulations they used were found to be compatible with white noise for these time units in relevant spatial and seasonal domains. Nevertheless, the knowledge that simulated unforced temperature variability can show autocorrelation motivates an extension of the SUN12 framework. Here we extend the theory by allowing unforced simulated temperatures to follow a short memory time series model, in particular AR(1). It is shown in Appendix A how this is achieved with simple adjustment factors for the variances of the R and T statistics, including a discussion how it is affected by the choice of time unit.

The second assumption must be relaxed in order to study the influence of two or more forcings added sequentially to a climate model, or to compare simulations with forcings of a similar type to see if one fits significantly better than the other. Sequentially included forcings has been implemented e.g. by Phipps et al. (2013), but is not satisfied by the Jungclaus et al. (2010) set of simulations. However, we want to use this data set to compare simulations driven by low- or high-amplitude solar forcings, and we demonstrate in Appendix B that this can be done by a significance test allowing a particular forcing to have influence on the real climate. The method is easily described. We simply calculate the standard error of the T statistic as if both simulations were unforced, and we will be on the safe side. The correlation test, on the other hand, must be changed, such that we compare the two R statistics with each other and not with zero.

For the question how data from several areas and/or seasons should be combined into a single test statistic, SUN12 (Sect. 7) proposed the use of a linear combination of the corresponding T-statistics. This requires an expression for the variance/covariance matrix of the set of T-statistics, based on the individual area/season standard errors and correlations of the T statistics, Eq. (1), leading to the final performance metric, U_T, for each climate model under consideration:

$$U_T = \frac{\sum_j c_j T_j}{\sqrt{\mathrm{Var}(\sum_j c_j T_j)}}.$$

Here, index j represents the various sites (areas/seasons) and the coefficients c_j indicate arbitrary weights that can be given to the sites. The denominator is the standard error of the numerator, and the test statistic U_T is approximately $N(0,1)$-distributed under the null hypothesis H_0 that the forcing introduced has no systematic effect on the fit of the model for any site.

If the forcing works realistically, we expect to see negative observed T-values and U_T-value, but if the model exaggerates the forcing effect, we might see systematically positive values. If a forced simulation produces a result that is indistinguishable from an unforced simulation (or a forced reference model, as in Appendix B), we would expect to see statistically insignificant T- and U_T-values, around zero. The correlation statistics $R(x,z)$ can be combined in the same way as the T statistics, into an aggregated correlation test value U_R (see SUN12, Sect. 8).

Before the statistical framework can be applied, the time resolution (time unit) to use for the model-data comparison must be decided. For reasonably correct test p values, it is essential to select a time unit that does not seriously violate an assumption that the simulated temperature for the reference model is AR(1). It is also necessary to select the size and shape of the area that a certain temperature (τ and x) represents. If areas of very different sizes are combined in the calculation of U_R and U_T, it may be motivated to choose appropriately different weights c_j. Different statistical precision of the proxy data series (z), however, does *not* motivate choosing different weights

c_j, because such differences are already accounted for by the weights w used in D^2_w and R. Data from different regions need not represent the same season. Regions may overlap and it is even possible to include data from different seasons for one and the same region. Proxy series from different regions may have different lengths. SUN12 proposed to achieve this by letting the number of time steps, n, be the same for all regions. Regions with shorter proxy records than the full analysis period will thus have no terms contributing to their D^2_w sums in periods when they have no data. This would be the same as having a proxy z with zero correlation to the true temperature τ, and thus a weight $w_j = 0$ before the actual proxy record starts. Evidently, several decisions need to be made when applying the framework in practice. Some of these issues will be discussed in Sect. 4, while Sect. 3 explains and discusses the choice of data sets.

3 Data

3.1 Climate model data

We follow Hind et al. (2012) and Hind and Moberg (2013) and use the simulations by Jungclaus et al. (2010) made with the Max Plack Institute Earth System Model (MPI-ESM)[1]. This comprises an atmospheric model run at T31 (3.75°) resolution and an ocean model run at a horizontal resolution varying between 22 km and 350 km. The MPI-ESM includes an interactive carbon cycle model comprising an ocean biogeochemistry model and a land surface scheme.

Jungclaus et al. (2010) performed several simulations with forcing histories starting at 800 CE and a 3000 year long unforced control experiment with orbital conditions as of 800 CE and constant pre-industrial greenhouse gas levels. Forced simulations of two kinds were made; one set with only a single forcing (either solar, or volcanic, or land cover change) and another set with multiple forcings (combining solar, volcanic

[1] http://www.ncdc.noaa.gov//paleo/metadata/noaa-model-10477.html

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

and land cover with orbital and greenhouse gas forcing as well as with non-volcanic aerosols). Two alternative solar forcing histories were used; the small-amplitude one by Krivova et al. (2007) with a 0.1 % change in total solar irradiance between the Maunder Minimum and the present, and the large-amplitude one by Bard et al. (2000) having a 0.25 % change. These two solar forcing series are not simply differently scaled versions of the same basic time series, but their long-term evolutions have rather similar character. The multiple-forcing simulations are available as two small ensembles, where individual members start from different ocean initial conditions at 800 CE. The "E1" ensemble, using the small-amplitude solar forcing, has five members while the "E2" ensemble, using the large-amplitude solar forcing, has three members. Like Hind et al. (2012) and Hind and Moberg (2013), we use forced simulations (x) from year 1000 CE onwards and split the control simulation into three 1000 year long segments to obtain a small ensemble of unforced simulations (x^*) of the same length. We refer to Figs. 1 and 2 in Hind et al. (2012) for time series plots of all forcings and of simulated global mean land-only temperatures for the various simulations by Jungclaus et al. (2010).

3.2 Instrumental data

Instrumental temperature data are needed for two purposes. First, SUN12 argued for using as good data as possible to maximize the statistical precision of the model-data comparison. Thus, in most cases, instrumental data should be preferred over proxy data within time periods when both exist. Second, instrumental data are needed to calibrate the proxy data. There are, however, several alternative temperature data sets to choose between (e.g. Brohan et al., 2006; Smith et al., 2008; Hansen et al., 2010; Morice et al., 2012).

Hind and Moberg (2013) used the CRUTEM3 land-only dataset by Brohan et al. (2006), having a 5° resolution going back to 1850. This dataset is provided with estimates of the error term in grid-box or larger-scale mean temperatures. However, Hind and Moberg (2013) could only incorporate these estimates in the SUN12 framework with assistance from the main author of Brohan et al. (2006), as error terms were not published for arbitrary regions and seasons. The updated land-plus-marine dataset HadCRUT4 (Morice et al., 2012) is provided with more comprehensive quantitative information about various types of errors, partly dealt with by presenting grid-point temperatures as 100 slightly different ensemble members. This should make it possible to estimate relevant noise terms, although at the expense of extra programming. We have not tried this option here.

For the current study, we selected instead the GISS1200 gridded global temperature dataset (Hansen et al., 2010) which goes back to 1880. This dataset uses a rather large search radius (1200 km) for averaging data from temperature stations in the calculation of each grid-point value. Therefore, GISS1200 data is spatially and temporally rather complete in remote areas such as the North American and Eurasian sub-Arctic regions, where several tree-ring chronologies are located (see below) but where few temperature stations – often with rather short records – are found. Despite its rather coarse spatial smoothing, GISS1200 is published at a rather fine grid (2°). This gives some flexibility when defining regions for temperature averages against which the tree-ring records are calibrated. Because the model and instrumental grids are different, we re-gridded the model grid to the same as for GISS1200 using bi-linear interpolation to enable comparison of analogous regions. A drawback with using GISS1200 is that explicit information about the instrumental error term is not available. We have therefore simply subjectively assumed that the noise term always accounts for 5 % of the total variance in instrumental temperature data, regardless of season and size of region. This is a limitation, but we checked the sensitivity of our results to the instrumental noise assumption by trying also 0 %, 10 % and 20 % (these results are not shown). This had only a marginal effect and did not affect any conclusions.

3.3 Tree-ring data

Tree-ring data are available from many parts of the globe. They can be sensitive to climate in different seasons but always have annual resolution and often explain a substantial fraction of observed temperature or precipitation variation (Fritts, 1976; Hughes, 2002; Briffa et al., 2004; Hughes et al., 2011; St. George and Ault, 2014). Tree-ring data, from either ring-width (TRW) or maximum density (MXD), are also the most extensively used proxies in temperature reconstructions for the last millennium (Jones et al., 2009). Here, we select fifteen tree-ring records that start before 1500 CE and which have been demonstrated to show a signal of temperature variability for a certain seasonal window. Four records are from North America, five from Europe, four from Asia and two from Oceania. Nine records start before 1000 CE, i.e. they extend back to the start of our analysis period. Table 1 lists all records with their short names used here, data type (TRW or MXD), seasonal targets, first year used in analysis and references to literature that describe the records. Table 2 provides web links to data source files. We regard our selection as sufficently complete for the purpose of this study although there are, admittedly, other records that could potentially have also been included. It is no problem that the SH and NH seasons are offset by half a year, because each site contributes its own R- and T-value to the U-statistics (see Sect. 2).

Twelve of the fifteen tree-ring records have been developed using the Regional Curve Standardization (RCS) technique (cf. Briffa et al., 1992), which can preserve variations on time scales longer than the life-length of individual trees. This is essential here, as we are interested in studying long-term temperature variations, in particular to distinguish between small- and large-amplitude solar forcing simulations. "Individual standardization" (IND) will inevitably inhibit variations on longer time-scales, as has been frequently discussed as the "segment length curse" problem in dendroclimatology (e.g. Cook et al., 1995). In fact, all standardization methods, whether applied as IND or RCS, will effectively remove a portion of the climate signal from the raw tree-ring data. To remedy this, Melvin and Briffa (2008) proposed a "signal-free" standardization method (SF), where an iterative procedure is used to separate the climate signal from the raw data. This procedure can be applied to both IND and RCS standardization, but very few records have been created with this rather new technique. Three records in our collection were developed using SF in combination with RCS.

One additional comment should be made in context of the SUN12 framework. The number of trees used in a tree-ring chronology will most often vary through time; typically there are fewer trees in the earliest part of a chronology, but the sample size can vary very irregularly with time. These variations in sample size are known to cause temporal variations in the variance of a chronology. Osborn et al. (1997) proposed to adjust the chronology variance such that it is approximately the same at each time point as if, hypothetically, an infinite number of trees from within the actual region had been used. This type of variance adjustment is nowadays a standard procedure in dendroclimatology and several records in our selection are processed this way. A somewhat similar variance adjustment is sometimes also applied to account for a varying number of chronologies used to build a composite temperature reconstruction, as for example in the records of Wilson et al. (2007) and Cook et al. (2013) used here. It may be that these variance adjustments induce a violation to a crucial assumption in the SUN12 framework, namely that a proxy sequence z should be calibrated such that the true temperature component τ always has its correct variance, with the noise term ϵ superimposed. Un-doing these adjustments is generally not possible without information that is only available to the original investigator, and it is beyond the scope of this study to attempt doing this. We merely point out this issue as a potential problem and simply regard published chronologies or temperature reconstructions as uncalibrated proxy sequences, which can be re-calibrated back to the start of the analysis period by using the statistical relationship to selected instrumental temperature data in a chosen calibration period.

4 Practical considerations

4.1 Selecting seasons

A first decision is to select the season that each proxy record will represent in the model-data comparison. As each original author team has generally spent consider-able efforts on determining the most appropriate season for each record – and as the SUN12 framework admits using all possible combinations of seasons – it seems most natural to follow the respective original judgements (see Table 1).

4.2 Choosing calibration periods

A time period (or time periods) is required for calibration of tree-ring data and, as we argue below (Sect. 4.3), for analysing the spatial pattern of correlations between tree-ring data and the instrumental temperature field. The longest possible period of overlap between instrumental and proxy data would maximize the number of observations used. However, in some cases proxy-data investigators have argued that either the tree-ring records are unreliable after a certain year (e.g. due to some local man-made disturbance; Cook et al., 2002) or the instrumental data appear to have lower quality before some point of time (Cook et al., 2013). Thus, our general recommendation is to use the longest meaningful calibration period for each record, but avoid using calibration data that are known to be unrepresentative. For the current study, however, we take a simple pragmatic approach and use the same calibration periods as were used by each original investigator (see Table 3).

4.3 Defining regions

Defining the region that each tree-ring series will represent is a more challenging task. SUN12 stated (in their Sect. 2), that “typically, this region consists of a single grid box, but averages over several grid boxes can also be considered”. A single grid-box temperature may perhaps maximize the statistical precision for calibration of a single tree-ring chronology, but a climate model can hardly simulate climate variability realistically within a single grid box. Also, one of our tree-ring records (ASIA2K) is derived from trees that grew in an area that extends over several grid boxes. Moreover, unforced temperature variability will have a larger influence in a single grid box as compared to an average of several grid boxes. Thus, as we are here primarily interested in seeing how well the model simulates the *externally forced* temperature variation, it appears recommendable to select an area that is large enough to detect the forced simulated temperature response, but small enough that the actual proxy record provides a meaningful approximation of the true temperature variability.

Although we cannot give a more precise recommendation, we can at least suggest a practically affordable way to semi-subjectively define a reasonable region for each tree-ring record. To this end, we plot and visually interpret the spatial field of correlations between each tree-ring record and the appropriate seasonal mean temperatures in GISS1200 data (Fig. 1). This correlation analysis is made using first-differenced data, to minimize possibly spurious correlations due to trends that do not reflect a direct physiological association between the temperatures in each growth season and the tree-ring data. This idea is similar to that adopted by Cook et al. (2013), in their screening to determine which individual tree-ring chronologies were positively correlated with grid-point temperatures, although they fitted an AR(1) model and removed this component from the data before calculating correlations.

The spatial correlation analysis was undertaken for calibration periods chosen above. Each map was then visually inspected to determine an appropriate region. We did not attempt to define any objective criterion, but we combined information about (i) where correlations are strongest, (ii) where chronologies are located and (iii) information from the literature regarding which regions the data represent. For example, the TASM area is allowed to extend over much of the ocean surrounding Tasmania, because Cook et al. (2000) suggested their record as a proxy for large-scale sea surface temperature anomalies. As another example, we followed the observation by Cook et al. (2013) that the ASIA2K record best represents regional temperatures north of 36° N. An additional constraint was the spatial resolution of the instrumental temperature grid (2°) and an account for the land/sea mask in the climate model and how this relates to the real land/sea borders. We did not attempt to merge all these pieces of information objectively, but our approach is merely an “expert judgement”. The resulting regional representation for each tree-ring record is illustrated in Fig. 2 and the regional latitude/longitude boundaries as well as corresponding fractions of the global area are provided in Table 3. Together, the fifteen regions represent 5 % of the global area but their sizes differ remarkably. The largest region (ASIA2K) is 70 times larger than the smallest (ALPS) and alone comprises more than 40 % of the total area of all regions put together.

4.4 Selecting weights c_j

The vastly different sizes of regions, as well as their uneven geographical distribution and different seasonal representation, motivates that some suitable weights c_j (explained in Sect. 2) are chosen. The simplest choice is to let all c_j be equal, i.e. to regard all selected proxy records as equally important. Another intuitive choice is to use the area of each region as weight. A third alternative is to choose weights according to how much "new" or "additional" information that each region contributes with in comparison with the other regions. A fourth alternative could be to weight the regions by how easy it is to detect the externally forced variability. Here, we try the first three alternatives and compare the results to see how sensitive U_R and U_T measures are to the choice of weights c_j. Moreover, we study the effect of excluding the three tree-ring records that were not RCS-standardized (GOA, CT, ASIA2K) and weighted the remaining twelve regions equally.

The equal and area weights are straightforward. The latter are provided in Table 3. Note that the sum of c_j need not be one. For the third alternative, we try a cluster analysis approach. This, however, requires some subjective decisions; one needs to choose a distance metric, a linkage method and also decide how many clusters to use. One also needs to decide which data to analyze. The full 3000 year control simulation is an adequate choice that provides a large sample representing unforced (internal) simulated climate variability. The quantity $1-r$, where r is the sample correlation between regions, appears intuitively meaningful as distance metric and Fig. 3 shows the result of a cluster analysis using nearest neighbour linkage (Matlab, 2008). By choosing seven clusters, we obtain a geographically and climatologically meaningful grouping of regions: Northern Scandinavia (JAMT, TORN), Continental Europe (PYR, ALPS, TATRA), Eastern Asia (ASIA2K), Oceania (TASM, NZ), Northwestern Siberia (AVAMT, YAMC), Northwestern North America (GOA, FIRTH, CT, CANR), Northeastern Siberia (YAK). We set the weights c_j such that each cluster contributes with one seventh to the total. Within each cluster, the contributing regions are equally weighted. This gives cluster-based weights as listed in Table 3.

4.5 Calibration of the tree-ring records

The tree-ring data need to be re-calibrated to appropriate regional and seasonal mean temperatures. Thus, the GISS1200 seasonal mean temperatures are averaged within each region and calibration is made for the chosen calibration periods, following procedures explained in Sect. 4 of SUN12 under the assumption that instrumental noise variance accounts for 5 % of the total observed temperature variance in each region (see Sect. 3.2). Moreover, as explained in Sect. 3.3 here, we assume that the statistical precision of each tree-ring record in the calibration period is also representative back to the start of the record. Table 3 lists correlations between each tree-ring record and the corresponding instrumental temperature record, ranging from 0.42 to 0.79. These correlations provide the information on the statistical precision of proxies that is used when calculating weights w. For each region, the calibrated tree-ring data sequence is then taken as the z sequence to compare with the corresponding model sequence x. The variance contribution from the calibration uncertainty has not been considered in our analysis.

4.6 Selecting analysis period

Another decision concerns the time window for which U_R and U_T measures are computed. With our choice of data, the longest possible window would be 1000–2000 CE, which includes both pre-industrial conditions and the increasingly anthropogenically influenced industrial period. Our focus, however, is on natural forcings, which motivates exclusion of the industrial period. We choose to analyse the period 1000–1849 CE to make it possible to directly compare our results with those from Hind and Moberg (2013). Thus, z sequences in our model-data comparison do not include any instrumental data. If we had chosen to include data after 1880, we would have used GISS data after 1880 and re-calibrated tree-ring data before 1880 (see Sect. 2).

4.7 Selecting time unit

Finally, a time unit must be selected; i.e. the length of time periods over which we average temperatures to obtain the pairs of simulation (x_i) and observation (z_i) values to be compared. It should be noted, though, that the precise choice of time unit is not crucial. Empirically, this can be seen in Fig. 6. Regarded as a question of principle, we must compromise between arguments for longer and shorter units of time. Arguments for long units are a reduced autocorrelation in the reference simulation and, provided there is little variation in the externally forced temperature component of x or z and in the weight w within units, a partial efficiency gain analogous to the gain by inside vs. outside averaging mentioned in Sect. 2. Arguments for short units are the anticipated within-unit variation in the forced component and (sometimes) in w, together with the need to estimate sample variances (see Sect. 5 in SUN12). The latter can be problematic, in particular, because the length of the available instrumental record poses an upper limit on how long time units that can be used. For example, with 120 years of instrumental observations, only four samples are present for estimation of the instrumental temperature variance if the chosen time unit would be 30 years. We have aimed at making time units short while controlling the autocorrelation.

The shortest possible time unit is dictated by the resolution of tree-ring data, which is 1 year. Thus, letting the time unit be 1 year would maximize the sample size. Therefore, we have always used the 1 year unit for calibration of the tree-ring records (in Sect. 4.5). However, before calculating U_R and U_T statistics, we need to check that the choice of time unit there will not seriously violate the assumption that unforced temperature variability can be approximated by an AR(1) process (see Appendix A).

To determine this, we analyze the autocorrelation in the 3000 year control simulation in two ways for each region. First, the lag-1 autocorrelation is computed for all time units from 1 to 30. Then, for a few selected time units (1, 3, 5, 8, 12 years), the autocorrelation function is estimated for lags up to 30. Figure 4 suggests that the lag-1 autocorrelation is in agreement with white noise except in some regions at short time units. Further, Fig. 5 (top) reveals that an AR(1) process is not sufficient at the 1 year unit within four regions (GOA, ASIA2K, TASM, NZ), which show a clear oscillatory behaviour with a period of about 3 to 4 years. As these four regions are located near the Pacific Ocean, a reasonable guess is that the model's El Niño/Southern Oscillation could be the cause. For the other four selected time units (Fig. 5, middle and bottom), we find support for either a white noise or an AR(1) assumption. Thus, for this study, we choose time units of 3, 5, 8 and 12 years to compute U_R and U_T statistics and compare the results. We use the AR(1) adjustment from Appendix A whenever the estimated lag-1 autocorrelation is positive. Although negative lag-1 autocorrelations may be physically meaningful in some cases (see e.g. Vyushin et al., 2012) we assume white noise is reasonable whenever estimated values are negative. Hence, we are on the safe side since negative lag 1 correlation is associated with a reduced standard deviation of the T-statistics. Moreover, when comparing two forced simulations directly, as in Appendix B, we also need to check that the forced simulated temperatures do not violate the AR(1) assumption. Thus, as we use this new approach to compare the E1 and E2 simulations, we made the same checks also for those data (results are not shown). As expected, we found evidence for stronger lag-1 autocorrelation than in the unforced control simulation, but an AR(1) assumption is valid in the large majority of cases. A few regions in E2 showed more persistence than expected from AR(1), but this does not affect our results since only E1 data are used to estimate the autocorrelation (see Appendix B4).

5 Results and conclusions from calculation of U_R and U_T statistics

Figure 6 shows calculated U_R and U_T statistics for individual regions and when all regions are combined in different ways, for the four selected time units. Notably, none of the single-forcing simulations robustly show U_R values above the 5 % significance threshold. Only the volcanic simulation shows some (barely) significant U_R values for the combined regions, but only for one or two time units. Despite this lack of significance, the multiple-forced E1 and E2 simulation ensembles show significant U_R values for all four time units and for all four regional weightings – actually with p values much smaller than 5 %. How could these highly significant values be explained? There are two main reasons. First, the forced component in simulated temperatures will stand out more clearly in an ensemble average than in a single simulation – in particular so for the larger E1 ensemble size. Second, the multiple-forced simulations do not only include volcanic, solar and land-use forcing, but also greenhouse-gas and orbital forcing. We can unfortunately not calculate any U_R values for the latter two forcings as no corresponding single-forcing simulations are available. Thus, we cannot judge how much these additional forcings contribute to the highly significant U_R values for the E1 and E2 ensembles. We can anyway conclude that both multiple-forced ensembles explain a statistically highly significant proportion of the temporal variation seen in tree-ring data. Thus, it is a meaningful exercise to see if they also fit the observations better than unforced simulations. This is done by studying their U_T values.

It turns out that U_T values for the combined regions are always negative for the E1 and E2 simulation ensembles (i.e. plotted above the zero line in Fig. 6). Thus, the multiple-forced simulations show smaller calculated D_w^2 distances to the tree-ring based observations than if unforced simulations are used. However, U_T values are not always significant at the 5 % level – but they are significant for some time units, or regional weightings, for both E1 and E2. So, the multiple-forced simulations are closer to the observations than unforced simulations, but it would be a too strong conclusion to say that they are "significantly closer".

Figure 7 is an attempt to graphically illustrate how well the E1 and E2 simulation time series match the tree-ring based observations and how they compare with the unforced control simulation. Although calculation of U_R and U_T values is made separately for each region, the figure for simplicity shows data averaged over all regions (and only for the 12 year unit). By eye, one can see a correlation between proxy data and the E1 and E2 simulations, but it is not easy to see if these forced simulations are "much better" than the control simulations – in intuitive agreement with the U_R and U_T results. It is also not easy to see if E2 is "better" than E1 or vice versa. But this can be tested by using the relaxed assumption in Appendix B, which permits the computation of U_T to test directly whether one of the E1 or E2 simulation ensembles is significantly closer to the proxy data than the other.

Figure 8 shows these U_T values for all four time units and for both outside and inside averaging. Clearly, results are highly dependent on which region is used. For some regions, E1 is "significantly better" than E2 but for some other regions the opposite is seen. For most regions, however, none is significantly better than the other. Unsurprisingly, no combination of all fifteen regions have significant U_T values at the 5 % level. Moreover, some weightings cause U_T values to be negative (support for E2; with large solar amplitude) while others show the opposite (support for E1; with small solar amplitude). The conclusion must be that, given the set of tree-ring data used here, it is not possible to tell which of the two multiple-forced simulation ensembles that best matches the observed temperature variations. As concerns the effect of including or excluding the three tree-ring records where RCS was not used, the result is somewhat in favour of the E2 simulation for the remaining twelve sites combined. Looking closely at details, however, this last result is much influenced by the non-RCS GOA record, where E1 gives a considerably better match.

6 Final discussion and conclusions

Practical application of the SUN12 framework (Sundberg et al., 2012) and certainly also other methods for paleoclimate model-data comparison, e.g. in data assimilation or detection and attribution studies, involve several decisions to be made by the investigator. A possible solution to handle this situation is to make a few alternative decisions and study how sensitive the results are. This approach is relevant in the current study, which is mainly methodological in nature. We studied the effect particularly related to two decisions; the choice of weighting information from different regions and the choice of time unit (time resolution). The latter was facilitated by an improvement of the framework to allow unforced simulated temperatures to follow an AR(1) process, rather than just white noise as in SUN12. This made it possible to choose time units down to 3 years, which is considerably shorter than the 20 or 30 years used in earlier studies by Hind and Moberg (2013) and Hind et al. (2012). Although an AR(1) assumption was empirically found valid for climate model data at time units used here, it could be motivated with further development of the framework to account also for the possibility that simulated climate shows stronger persistence, such as a power law, as has been found by e.g. Vyushin et al. (2012).

In this study, we used an ensemble of climate model simulations run with forcing conditions for the last millennium (Jungclaus et al., 2010), which we compared with a set of fifteen tree-ring based temperature proxy data series representing regions of different size, different seasonal mean temperatures and having different lengths and statistical precision. Our results showed that, among the single-forcing simulations (land-use, small-amplitude solar, large-amplitude solar, volcanic), only the one with volcanic forcing could with statistical significance explain any of the observed variations in the pre-industrial period 1000–1849 CE – but only for one or two of four time units tried depending on which regional weighting was used. This finding, that only the effect of volcanic forcing – but not solar forcing – could be significantly detectable in proxy data is in agreement with results from detection and attribution studies both at a hemispheric scale (Hegerl et al., 2007) and a European scale (Hegerl et al., 2011).

When all forcings were combined (land-use, small-amplitude *or* large-amplitude solar, volcanic, orbital, greenhouse-gas) and also used in small simulation ensembles, the simulations were however highly able to capture some of the observed temperatures as recorded in tree-ring data. We cannot identify the precise reason(s) behind the significant test values, but it should be partly related to inclusion of orbital and greenhouse-gas forcings and partly because the response to forcings is expected to stand out more clearly in an ensemble average than in a single simulation, as has been demonstrated in pseudoproxy experiments (Hind et al., 2012). Both multiple-forced simulation ensembles are closer to the tree-ring based observations than if unforced control simulations are used, implying that the temperature response to the combination of forcings is realistic. However, results at the individual regional level differ greatly and significance is not reached for all time units and choices of weights when regions are weighted together. A conclusion here is that an average of many sites is needed, or the separate sites/records need to be more clearly classified as less or more reliable and representative than others.

Another improvement to the SUN12 framework made it possible to test directly if one of the two multiple-forced simulation ensembles (i.e. including *either* small- or large-amplitude solar forcing) is closer to the observed temperature variations than the other. However, results were highly dependent on information at the regional level, which made it impossible to judge if any simulation ensemble is “better” than the other. Thus, this new analysis based only on tree-ring data from several regions did not show any clearer results than a previous northern hemispheric-scale study based on several compilations of different proxy data (Hind and Moberg, 2013). This inconclusiveness is perhaps not surprising, given that differences between simulations with frequently used weaker or stronger solar forcing are rather small (Masson-Delmotte et al., 2013). The new results, however, give some weak support for the large-amplitude solar forcing, whereas the Hind and Moberg (2013) study pointed in the opposite direction. Although the weaker solar forcing is more in line with most recent viewpoints (Masson-Delmotte et al., 2013), it is still possible that none of the two alternative solar forcings is correct and that the truth is somewhere in between. An extension of the framework to allow estimation of how well the amplitude of a true external forcing is represented in a simulation could help to provide a more informative answer. As already argued in SUN12 (Sect. 9), such an extension would also bring their framework closer to that used in detection and attribution studies (e.g. Hegerl et al., 2007; Schurer et al., 2013).

One may ask as to what extent the choice of using only tree-ring data has influenced the results. For example, their inability to correctly capture the long-term trend on millennial scales has been discussed by Esper et al. (2012). This problem should be most prominent in records where RCS standardization was not used. Omitting the three non-RCS records in our collection had the effect to give somewhat more support for the large-solar amplitude forcing, although this was actually due to the property of only one of the three non-RCS records. We have also argued that variance stabilization procedures (Osborn et al., 1997) applied to many tree-ring chronologies are in conflict with assumptions in the SUN12 framework. This may affect results, presumably to make statistical test values “too significant”. Another potential problem is the observed spectral biases in many tree-ring records (they are often too “red”; Franke et al., 2013). This does not affect the validity of the tests, but will affect their power. It remains to analyze how these and other problems, e.g. regarding the different nature of response in TRW and MXD data to volcanic forcing (Esper et al., 2013; Jones et al., 2013), affect results from model vs. tree-ring data comparisons.

Information also from other types of proxy data should potentially help to more conclusively compare a set of alternative simulations with proxy-based climate observations. All our proxy records reflect temperatures only in the tree-growth season; i.e. mostly a summer or an extended summer season. Perhaps the regions used here are too small, or too few, or not sufficiently well distributed in space, and in combination with a lack of information from winter this might cause internal unforced variability to dominate too much over the response to external forcings? A model study by Servonnat et al. (2010) suggested that the response to external forcings are only detectable within regions larger than approximately the size of Europe, thus pointing to the importance of not using too small regions in studies like this. On the other hand, the pseudoproxy study by Hind et al. (2012) suggested that annual-mean temperature data, with realistic proxy noise levels, from at least 40 randomly distributed single grid-boxes are needed to clearly separate between the two sets of multiple-forcings used here. Thus, averaging information from a sufficient number of small regions can be meaningful, even if each region by itself is too small to clearly separate the externally forced signal from internal climate variability.

There are certainly many more published proxy records (and more are expected to appear in the future) that could potentially be used in this type of model-data comparison studies. But there is still somewhat open regarding whether proxy data are best used as individual records, as most records in this study, or aggregated into larger-scale averages such as in the PAGES2K dataset (PAGES2k Network, 2013). In that case, seven continental-scale annual-mean or summer-mean temperature reconstructions (including ASIA2K used here) were derived from different types of proxy data. This latter approach has the potential advantage of reducing the influence from various types of noises, both in proxy data and from internal variability in both models and real climate. A drawback, though, is that seasonally specific information in each proxy is partially lost and the optimal region and season for each large-scale data aggregate is essentially unknown. Thus, more theoretical and practical work addressing questions such as the optimal spatial analysis scale is motivated – in parallel with continued development of climate models, forcing datasets and climate proxy records.

Appendix A: Adjustment for autocorrelation in the reference climate model simulation series

This is a derivation of an adjustment factor for the correlation test statistic R and for the D^2-based test statistic, necessary to allow autocorrelation in the reference climate model simulation series, in particular under an AR(1) model for this autocorrelation. Finally, an MA(1) model and the effects of a k years time unit are also treated.

A1 Model

Suppose we have a climate model[2] x with time-varying forcing, and another, x^*, being a reference free from such forcings. We also have an observation series for the same period as the forced model, denoted z. The observations z_i (time step $i = 1, \ldots, n$) represent instrumental measurements when such are available, otherwise a proxy assumed to be correctly calibrated. We want to test first if the forced model x shows evidence of a correlation with the observations (R-test), and next if it fits the observations better than the reference model x^* (D^2-based test). Of concern here is the performance of the test statistics when there is autocorrelation present in both climate models. As hypothesis model, we take x and x^* to be mutually equivalent and autocorrelated AR(1), but uncorrelated with the true and measured temperatures, τ and z:

Statistical Model under H_0: Climate model simulation sequences $\{x_i\}$ and $\{x_i^*\}$, true climate sequence $\{\tau_i\}$, and observation sequence $\{z_i\}$ are mutually related through the following model:

$$\begin{aligned} x_i &= \mu_x + \delta_i, \quad \mathrm{Corr}(\delta_i, \delta_{i-k}) = \rho_k \\ x_i^* &= \mu_x + \delta_i^*, \quad \mathrm{Corr}(\delta_i^*, \delta_{i-k}^*) = \rho_k \\ \tau_i &= \mu_\tau + \eta_i, \\ z_i &= \tau_i + \epsilon_i. \end{aligned}$$

[2] With "climate model", we think of a realization of an Atmosphere-Ocean General Circulation model or an Earth System Model integrated in time, with or without time-varying external forcings. The variable x represents simulated temperatures in a certain region and season of interest.

Note that there is no forcing effect in this model, but the test statistics were designed to be sensitive to a forcing effect in common for x and τ. The variates x_i and x_i^* have the same mean value and mutually uncorrelated "noise terms" δ_i and δ_i^*. Other terms representing unexplained variability (random fluctuations, internal variability, noise) are η_i and ϵ_i. Here η_i represents the true climate variability. We make no assumption about that variability. Technically, the observed z series is regarded as given and fixed, and the statistical analysis is conditional on this given series. Weights $w_i \geq 0$ and $\tilde{w}_i \geq 0$ (see Sects. 5 and 8 in SUN12) are also regarded as given and fixed.

We consider the correlation test and the D^2-based test, based on the same basic statistics as in the absence of autocorrelation, but we have to modify their variances (or standard errors) in order to allow autocorrelation. For the correlation test we do not need the unforced climate model, since the hypothesized correlation is known, being zero.

A2 Correlation test statistic

The weighted empirical regression coefficient $R(x,z)$ is used as test statistic, after normalization by its standard error, see Eqs. (19) and (20) in SUN12. Now, $R(x,z)$ differs only by a constant factor from $\sum \tilde{w}_i(x_i - \mu_x)z_i$. We need an expression for its variance, allowing some degree of autocorrelation.

First we note that since $\tilde{w}_i$ and z_i are both fixed and given, we may introduce a new weight factor $\dot{w}_i = \tilde{w}_i z_i$, being their product. Thus, we consider the variance of

$$\sum \dot{w}_i x_i$$

We start by the general formula for the variance,

$$\mathrm{Var}\left(\sum_i \dot{w}_i x_i\right) = \sigma_x^2 \sum_{i,j} \dot{w}_i \dot{w}_j \rho_{j-i} = \sigma_x^2 \left(\sum_i \dot{w}_i^2 + 2\sum_{i<j} \dot{w}_i \dot{w}_j \rho_{j-i}\right)$$

We now make the assumption that the x_i time series is an AR(1) process with lag 1 correlation ρ:

$$x_i - \mu_x = \rho(x_{i-1} - \mu_x) + \tilde{\delta}_i, \quad |\rho| < 1, \tag{A1}$$

where $\tilde{\delta}$ is a new error term. In such a model, the lag $j-i$ correlation ρ_{j-i} decreases exponentially with the time distance $j-i$, $\rho_{j-i} = \rho^{j-i}$. Below we also assume $\rho \geq 0$, which appears realistic if x_i is AR(1). We now get

$$\mathrm{Var}\left(\sum_i \dot{w}_i x_i\right) = \sigma_x^2 \left\{ \sum_i \dot{w}_i^2 + 2 \sum_{k=1}^{n-1} \rho^k \sum_{i>k} \dot{w}_i \dot{w}_{i-k} \right\} \tag{A2}$$

This exact value can be used, but we will also give a simple upper bound to it, that we have used in this paper. We use $\sum_{i>k} \dot{w}_i \dot{w}_{i-k} \leq \sum_i \dot{w}_i^2$, by Cauchy's formula, and insertion of this upper bound yields

$$\mathrm{Var}\left(\sum_i \dot{w}_i x_i\right) \leq \sigma_x^2 \sum_i \dot{w}_i^2 \left\{ 1 + 2 \sum_{k>1} \rho^k \right\} \leq \sigma_x^2 \sum_i \dot{w}_i^2 \frac{1+\rho}{1-\rho}$$

The last inequality is when a finite sum of ρ^k over k to $n-1$ is majorized by the corresponding infinite sum.

Thus we have an upper bound for the variance as a function of ρ. The variance factor

$$(1+\rho)/(1-\rho) \tag{A3}$$

is what differs from the case $\rho = 0$. For the standard error, we use its square root. For ρ small, the variance factor is about $1 + 2\rho$, or for the standard error $1 + \rho$, but this approximation is no longer an upper bound, so $(1+\rho)/(1-\rho)$ is preferable.

The inequality above when the finite sum was replaced by an infinite sum should typically be very close to an equality. On the other hand, the inequality motivated by

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

CPD

10, 2627–2683, 2014

Statistical framework for evaluation of climate model simulations – Part 3

A. Moberg et al.

Title Page

Abstract | Introduction

Conclusions | References

Tables | Figures

◂◂ | ▸▸

◂ | ▸

Back | Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Cauchy's formula is likely to be a large exaggeration of the actual value. There are two parts in this inequality. First, the sum of squares, $\sum_i \dot{w}_i^2$, contains n terms, whereas the sum of products, $\sum_{i>j} \dot{w}_i \dot{w}_{i-j}$, contains only $n-j$ terms. However, since the contributions from ρ^j with small j are likely to dominate this will make little difference. The second part concerns the magnitude of sums of products relative to the sum of squares. Here we must bring in the structure of the z series of real climate plus noise. A sum of products relates to the covariance of the z series, and if there is not very high autocorrelation in the z series, the sum of products will be much smaller than the sum of squares (representing the variance). Thus, multiplying the variance in Eq. (20) in SUN12 by the factor $(1+\rho)/(1-\rho)$ is likely to markedly exaggerate the effects of AR(1) autocorrelation in the x series. Nevertheless, this is how we made the adjustments in this study. When data from different regions are combined, one adjustment has to be calculated for each region. Then, the covariances in Eq. (21) in SUN12 are multiplied by the square root of the product of the pairs of adjustment factors.

A3 D^2 difference test statistic

The D^2 difference test statistic $T(x, x^*, z) = D_\mathrm{w}^2(x, z) - D_\mathrm{w}^2(x^*, z)$, see Eq. (1), can be expressed in the form

$$T(x, x^*, z) = \overline{w\,(x-\mu_x)^2} - \overline{w\,(x^*-\mu_x)^2} - 2\,\overline{w\,(x-x^*)\,(z-\mu_x)}, \tag{A4}$$

with over-line denoting averaging over time (n time steps). Here, when x and its reference x^* are equivalent, the first two terms have equal expected values and the third term will have the expected value zero, so T has expected value zero. Autocorrelation in x and x^* does not change this. To specify a test statistic we only need the variance of T in the hypothesis model.

The first two terms of Eq. (A4) are mutually uncorrelated. Each of them is also uncorrelated with the third term, under an assumption of Gaussian noise δ and δ^* in x and x^*, respectively (already made in SUN12). This is because the covariances will be proportional to the third order central moments of δ (or δ^*), which are zero because of the symmetry of the Gaussian distribution around its mean value (the only property needed, in fact). Note that $x - x^* = \delta - \delta^*$. Thus, all three terms are mutually uncorrelated, so we need only consider the sum of their respective variances.

The last term is linear in x, and it is the difference between two uncorrelated terms of the same type as the statistic studied in the previous section. By the same argumentation as there, we find that a safe variance adjustment factor is $(1+\rho)/(1-\rho)$, which is again likely to be an exaggerated adjustment.

The first two terms are of the same type, so we need only study one general such statistic,

$$\sum w_i (x_i - \mu_x)^2 .$$

Note that z is not involved here, so the weight factor is the more slowly varying w_i, not the $\dot{w}_i$ from Appendix A2. Under the AR(1) model (A1), we have

$$(x_i - \mu_x)^2 = \rho^2 (x_{i-1} - \mu_x)^2 + 2\rho\, (x_{i-1} - \mu_x)\, \tilde{\delta}_i + \tilde{\delta}_i^2$$

It follows that the covariance between $(x_i - \mu_x)^2$ and the corresponding preceding term is

$$\mathrm{Cov}\{(x_i - \mu_x)^2, (x_{i-1} - \mu_x)^2\} = \rho^2\, \mathrm{Var}\left((x_{i-1} - \mu_x)^2\right)$$

This is because $\tilde{\delta}_i$ and x_{i-1} are mutually independent. Continuing further steps back in time we get

$$\mathrm{Cov}\left\{(x_i - \mu_x)^2, (x_{i-k} - \mu_x)^2\right\} = \rho^{2k}\, \mathrm{Var}\left((x_{i-k} - \mu_x)^2\right).$$

Now we can do the same type of calculation as for the linear type of term above, and get the variance adjustment factor

$$\frac{1+\rho^2}{1-\rho^2} \tag{A5}$$

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

Note that the previous ρ has been replaced by ρ^2, which makes this adjustment factor closer to 1. Also, this adjustment factor is not likely to have an exaggerating influence, because the weight is now w_i, not $\dot{w}_i$. Since w_i will mostly change slowly with i, we have $\sum w_i w_{i-k} \approx \sum w_i^2$ for small k.

There are two possible strategies when choosing the adjustment factor. Either we simply use the formula in Eq. (A3) for the whole variance of T, which is then a deliberate over-adjustment, or we split T in its three components and use their respective variances with different adjustments for the different components. In this study, we have used the first (simpler) alternative. Thus, the variance in Eq. (15) in SUN12 has been multiplied by the factor in Eq. (A3). When data from different regions were combined, we calculated one adjustment for each region. Then, the covariances in Eq. (16) in SUN12 were multiplied by the square root of the product of the pairs of adjustment factors. In other words, we used the same adjustment for the correlation and the difference tests.

A4 Autocorrelation and time units

The results above were derived under an AR(1) model for the unforced climate simulations. Figure 5, top, shows the corresponding estimated autocorrelation functions for our annual data. Even if some regions appear consistent with the exponentially decreasing autocorrelation function of an AR(1), other regions show a damped sine wave type function, indicating an AR(2) process, or worse. The damping factor is of magnitude 0.85 year^{-1}. An AR(2) model would make the previous calculations considerably more complicated. Going further away from AR(1) by a model with long-range dependence will change the situation completely.

Hind et al. (2012) used a longer time unit to make all correlation negligible. With a not so long time unit, the lag 1 correlation is perhaps not negligible. However, because of the time gap between time units at lag ≥ 2 distance, correlations for lag 2 (or more) are likely to be much smaller than as prescribed by AR(1) (which is ρ^2). As a numerical example, suppose we have an AR(1) for annual data with $\rho = 0.45$. If we change time unit to 3 years, lag 1 correlation is of course reduced, but of interest here is that the lag 2 correlation is reduced much more, to be a factor $\rho^3 \approx 0.1$ times lower than the corresponding lag 1 correlation. More generally, for a series whose autocorrelations in the moderately long run decrease like in an AR(1) series, only a moderately long time unit is needed to make all lag ≥ 2 autocorrelations of the aggregated series negligible. Such a time series is represented by MA(1). Going through the derivations above, when there is only lag 1 correlation, it is seen that the adjustment factor is now closer to 1. More precisely, the denominator can be replaced by 1 in the formulas in Eqs. (A3) and (A5).

Whether MA(1) is a reasonable description must be judged from data. Figure 5 illustrates the situation. With time unit 3 years or 5 years we see a few significant lag 2 correlations of magnitude 0.2, but for longer time units the estimated lag ≥ 2 autocorrelations look like expected for white noise. Even with a damping factor 0.85 in an AR(1), we can conclude that the lag 2 correlation with an 8 years time unit is reduced relative to the lag 1 correlation by a factor $0.85^8 = 0.27$, and by one more such factor for lag 2, etc. Thus, if we use a time unit of 8 years and modify the test statistic variances by the factor in Eq. (A3), where ρ is the lag 1 correlation with this time unit, we should be on the safe side.

Appendix B: The test statistics in the presence of a joint forcing

In the situation that a particular forcing effect is present in the climate model simulations both with and without the forcing of interest, we show that the correlation test statistic R must be compared with the correlation for the reference model (null model), and that the D^2-based test statistic need not be adjusted at all. None of the forcing effects need be present in the true climate, but the tests discussed here are most likely of interest when the effect of the forcing of the reference model has already been detected in the observations or is assumed for physical reasons to affect the true climate. In

Appendix B4, we extend the situation to discuss comparison of two alternative forcings of the same type, such as low- and high-amplitude solar forcing.

B1 Model

As in Appendix A1, suppose we have two climate models[3], represented by simulation sequences x and x^*, the latter having a role as reference, and an observation series for the same period, denoted z. The new feature is that we allow a "baseline" forcing present in both climate models, and probably also in the true temperature. This baseline forcing is not of current interest, but there is another, additional forcing applied in x but not in x^*. As before, the hypothesis H_0 to be tested assumes this additional forcing has no effect.

Statistical Model under H_0: Climate model simulation sequences $\{x_i\}$ and $\{x_i^*\}$, true climate sequence $\{\tau_i\}$, and observation sequence $\{z_i\}$ are mutually related through the following model:

$$
\begin{aligned}
x_i &= \mu_x + \gamma_i + \delta_i \\
x_i^* &= \mu_x + \gamma_i + \delta_i^* \\
\tau_i &= \mu_\tau + \eta_i \\
z_i &= \tau_i + \epsilon_i
\end{aligned}
$$

Here, the term γ_i, in common for x and x^*, is the baseline forcing effect, regarded as of more or less random character. Thus, x and x^* differ only by separate random "noise" terms δ and δ^*. All three terms γ, δ and δ^* are assumed mutually uncorrelated with time-constant variances (even for γ when it is considered random, variance σ_γ^2). The baseline forcing may also be present in the true climate, and is even likely to be so. We therefore allow γ_i to be correlated with the term η_i, with no need to be more specific.

[3] See footnote in Appendix A1

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

In SUN12, the additional forcing of concern was represented by terms ξ and $\alpha\xi$ in the expressions for τ and x, respectively, but here we need not be so explicit because they do not occur in the null model. We always know, of course, that such an additional forcing has been implemented in the climate model simulation represented by x in the statistical model, but we want to investigate if a response to this forcing is also seen in the observations. Our tests are designed to detect if there is a strong enough such additional forcing effect jointly present in x and τ. If the test results lead to rejection of H_0, it indicates that there is an effect of the additional forcing in the true climate sequence τ, because our test statistics are sensitive only to a joint effect in x and τ.

For simplicity we assume here that there is no autocorrelation in the x-sequences. However, such autocorrelation can be adjusted for as described in Appendix A, and we did so in our experiment. Generally, this is likely to be even more needed here than before, since the baseline forcing effect γ is likely to contribute additional autocorrelation within the reference series.

Other terms representing unexplained variability (random fluctuations, noise) are η_i and ϵ_i, the η_i term representing all variability in the true climate. However, we make no assumption about the true climate τ or the observed climate series z. In particular, it may contain more or less effect from the baseline forcing that was also behind the γ term of the climate models. The reason we do not need assumptions is that we will consider statistics such as the difference between two D^2 values. The D^2 values themselves are typically reduced if we introduce a realistic forcing effect γ_i in the models, that is also present in the true climate. The difference between two such characteristics, however, will not be systematically changed. Technically, we regard the observed z series as given and fixed, and the statistical analysis is conditional on this given series. The weights $w_i \geq 0$ and $\tilde{w}_i \geq 0$ (see Sects. 5 and 8 in SUN12) will also be regarded as given and fixed.

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

B2 Correlation test statistic

The test statistic denoted $R(x,z)$ in Sect. 8 of SUN12 differs only by a constant factor from

$$\sum \tilde{w}_i(x_i-\mu_x)z_i. \tag{B1}$$

Here the theoretical average μ_x will be replaced by the corresponding empirical average.

When a forcing is present in the reference model, we must expect this forcing causes an underlying positive correlation with the observations on its own. For that reason we must bring in the reference x^* and show that x, as compared with x^*, is more correlated with z. Therefore we use the difference $R(x,z)-R(x^*,z)$ instead of $R(x,z)$. The γ term cancels, so given z, the difference consists of two mutually uncorrelated terms with variance twice the single term variance in Eq. (20) given in SUN12 as a function of σ_δ^2. It only remains to remember what σ_δ^2 stands for. This is the residual variance in the reference simulations x^* after adjustment for the unknown forcing effect γ. But this variance is majorized by the total variance of x^*, obtained when we additionally include the variation of γ_i, $\mathrm{Var}(x_i^*)=\sigma_\delta^2+\mathrm{Var}(\gamma_i)\geq\sigma_\delta^2$ if γ is regarded as random with a variance. When the γ_i sequence is regarded as fixed, we instead state that σ_δ^2 is overestimated by the total sample variance s_x^2 of the x^*-sequence. Thus we are on the safe side when using the sample variance of x_i^*, and in comparison with the results of SUN12 we need not bother about γ_i but simply adjust the variance in their Eq. (20) by a factor 2 when $R(x,z)$ is replaced by $R(x,z)-R(x^*,z)$. Furthermore, in many cases the relative difference between σ_δ^2 and s_x^2 will be small, so the majorization (upper bound) is not only on the safe side but also innocent.

B3 D^2 difference test statistic

The D^2 difference test statistic $T(x, x^*, z) = D^2_{\mathrm{w}}(x, z) - D^2_{\mathrm{w}}(x^*, z)$, see Eq. (1), can be expressed in the form

$$T(x, x^*, z) = \overline{w\,(x-\mu_x)^2} - \overline{w\,(x^*-\mu_x)^2} - 2\,\overline{w\,(x-x^*)\,(z-\mu_x)}, \tag{B2}$$

with over-line denoting averaging over time (n time steps). (Note: Eq. (B2) is identical to Eq. (A4).) Under H_0', saying that x and its reference x^* are equivalent, the first two terms have equal expected values and the third term will have the expected value zero, so T has expected value zero. This is true even when x and x^* have a term γ in common. To form a test statistic we only also need the variance of T under H_0'. Because it simplifies the derivation, we will here regard the γ term as random.

The first two terms of Eq. (B2) are mutually uncorrelated. Each of them is also uncorrelated with the third term, under an assumption of Gaussian noise δ and δ^* in x and x^*, respectively (already made in SUN12). This is because the covariances will be proportional to the third order central moments of δ (or δ^*), which are zero because of the symmetry of the Gaussian distribution around its mean value (the only property needed, in fact). Note that $x - x^* = \delta - \delta^*$.

The third term of Eq. (B2) yields a variance that is formally the same as in SUN12. The only difference is (again; cf. the previous section) in the interpretation of the unknown σ^2_δ. By using instead the sample variance of the reference x^*-sequence we get a useful upper bound.

The first two terms of Eq. (B2) have the same variance. In SUN12 this was given to be $2(\sigma^2_\delta)^2$, which was estimated by the sample variance of the reference model simulation. In the present case, when we consider γ_i as random (and Gaussian and uncorrelated with δ_i), we immediately get the same type of formula, but with $\mathrm{Var}(\gamma + \delta) = \sigma^2_\gamma + \sigma^2_\delta$ for σ^2_δ. In practice, there is no difference, however, because the natural estimate of this variance is still the sample variance of the x^*_i sequence.

We conclude that also for the first two terms of Eq. (B2) we can use the formula of SUN12, with its σ_δ^2 interpreted as the sample variance of the x^*-sequence. In other words, for the T-based test we can use the same calculation procedure as in SUN12, in particular their variance formulas in Eqs. (15) and (16), without bothering about γ_i, just pretending it does not exist. With the interpretation above, it does not matter what the γ_i sequence is. We have an upper bound for the variance, that will be close to the unknown true value unless the actual quadratic variation in the γ_i sequence is a substantial part of the total variance. To adjust for autocorrelation in the reference model, the lag 1 correlation ρ should be estimated from the sample x^* sequence, and a variance adjustment should be made as in Appendix A.

B4 Comparison of climate models with the same type of forcing

Additionally, the result above can be used to compare two climate models of the same kind, but driven with alternative versions of the type of forcing of interest, to see if one is significantly better than the other. We then test the hypothesis that the two models are equivalent, in the sense of having the same forcing and the same magnitude of the response to this forcing. Expressed in terms of the Statistical Model in Appendix B1 above, we test the hypothesis that the two simulation models have the same forcing effect term (the γ term), and if their D^2 difference is statistically significant we can conclude that one of the models fits better than the other. We do here as when we tested a forced model against an unforced control by forming a variance-normalized D^2 difference, although this test is now two-sided since none of the models is a reference. Thus we need the variance of the D^2 difference when both models have the same forcing effect, as in the previous section. A difference, though, is that the estimate of σ_δ^2 is now naturally taken to be the average of the sample variances for the two models.

One additional comment is motivated here. If the two forcings of interest are truly different alternatives with somewhat different temporal evolution, then, clearly, none of the models is a reference. But if the two forcings are just differently scaled versions of the same basic data, thus differing only in their amplitude, then the one with the smaller amplitude could be regarded as a reference, at least for the correlation test.

In our experiment where we compared the E1 and E2 simulations, the situation is somewhat in between, as the solar forcings differ both in low-frequency amplitude and in temporal evolution. Because the different amplitude is of the largest interest, we decided to estimate σ_δ^2 (and ρ) only from E1 (having the smaller solar forcing amplitude), but used a two-sided test for the result in Fig. 8. However, we also found that using an average of E1 and E2 parameter estimates hardly changed the results at all.

Acknowledgements. This research was funded by the Swedish Research Council (grants 90751501, B0334901 and C0592401; the first two given to the lead author, the third to Dr. Gudrun Brattström at the Department of Mathematics, Stockholm University).

References

Anchukaitis, K. J., D'Arrigo, R. D., Andreu-Hayles, L., Frank, D., Verstege, A., Curtis, A., Buckley, B. M., Jacoby, G. C., and Cook, E. R.: Tree-ring-reconstructed summer temperatures from northwestern North America during the last nine centuries, J. Climate, 26, 3001–3012, 2013. 2673

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. 2638

Briffa, K. R., Jones, P. D., Bartholin, T. S., Eckstein, D., H., S. F., Karlén, W., Zetterberg, P., and Eronen, M.: Fennoscandian summers from AD 500: temperature changes on short and long timescales, Clim. Dynam., 7, 111–119, 1992. 2640

Briffa, K. R., Osborn, T. J., and Schweingruber, F. H.: Large-scale temperature inferences from tree rings: a review, Global. Planet. Change, 40, 11–26, 2004. 2640

Briffa, K. R., Shishov, V. V., Melvin, T. M., Vaganov, E. A., Grudd, H., Hantemirov, R. M., Eronen, M., and Naurzbaev, M. M.: Trends in recent temperature and radial tree growth spanning 2000 years across northwest Eurasia, Philos. Trans. R. Soc. London, Ser. B, 353, 2269–2282, 2008. 2673

Briffa, K. R., Melvin, T. M., Osborn, T. J., Hantemirov, R. M., Kirdyanov, A. V., Mazepa, V. S., Shiyatov, S. G., and Esper, J.: Reassessing the evidence for tree-growth and inferred temperature change during the Common Era in Yamalia, northwest Siberia, Quaternary. Sci. Rev., 72, 83–107, 2013. 2673

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. Atmos., 111, D12106, doi:10.1029/2005JD006548, 2006. 2638, 2639

Büntgen, U., Frank, D. C., Nievergelt, D., and Esper, J.: Summer temperature variations in the European Alps, A.D. 755–2004, J. Climate, 19, 5606–5623, 2006. 2673

Büntgen, U., Kyncl, T., Ginzler, C., Jacks, D. S., Esper, J., Tegel, W., and Heussner, K.-U.: Filling the eastern European gap in millennium-long temperature reconstructions, PNAS, 110, 1773–1778, 2013. 2673

Cook, E. R., Briffa, K. R., Meko, D. M., Graybill, D. A., and Funkhouser, G.: The 'segment length curse' in long tree-ring chronology development for palaeoclimatic studies, Holocene, 5, 229–237, 1995. 2640

Cook, E. R., Buckley, M. M., D'Arrigo, R. D., and Peterson, M. J.: Warm-season temperatures since 1600 BC reconstructed from Tasmanian tree rings and their relationship to large-scale sea surface temperature anomalies, Clim. Dynam., 16, 79–91, 2000. 2643, 2673

Cook, E. R., Palmer, J. G., and D'Arrigo, R. D.: Evidence for a 'Medieval Warm Period' in a 1100 year tree-ring reconstruction of past austral summer temperatures in New Zealand, Geophys. Res. Lett., 29, 12-1–12-4, 2002. 2642, 2673

Cook, E. R., Buckley, B. M., Palmer, J. G., Fenwick, P., Peterson, M. J., Boswijk, G., and Fowler, A.: Millennia-long tree-ring records from Tasmania and New Zealand: a basis for modelling climate variability and forcing, past, present and future, J. Quaternary Sci., 21, 689–699, 2006. 2673

Cook, E. R., Krusic, P. J., Anchukaitis, K. J., Buckley, B. M., Nakatsuka, T., and Sano, M.: Tree-ring reconstructed summer temperature anomalies for temperate East Asia since 800 C.E., Clim. Dynam., 41, 2957–2972, 2013. 2641, 2642, 2643, 2673

D'Arrigo, R., Jacoby, G., Buckley, B., Sakulich, J., Frank, D., Wilson, R., Curtis, A., and Anchukaitis, K.: Tree growth and inferred temperature variability at the North American Arctic treeline, Global. Planet. Change, 65, 71–82, 2009. 2673

D'Arrigo, R. D., Wilson, R., and Jacoby, G.: On the long-term context for late twentieth century warming, J. Geophys. Res.-Atmos., 111, D03103, doi:10.1029/2005JD006352, 2006. 2673

Dorado Liñán, I., Büntgen, U., González-Rouco, F., Zorita, E., Montávez, J. P., Gómez-Navarro, J. J., Brunet, M., Heinrich, I., Helle, G., and Gutiérrez, E.: Estimating 750 years of temperature variations and uncertainties in the Pyrenees by tree-ring reconstructions and climate simulations, Clim. Past, 8, 919–933, doi:10.5194/cp-8-919-2012, 2012. 2673

Esper, J., Frank, D. C., Timonen, M., Zorita, E., Wilson, R. J. S., Luterbacher, J., Holzkamper, S., Fischer, N., Wagner, S., Nievergelt, D., Verstege, A., and Büntgen, U.: Orbital forcing of tree-ring data, Nat. Clim. Change, 2, 862–866, 2012. 2652

Esper, J., Schneider, L., Krusic, P. J., Luterbacher, J., Büntgen, U., Timonen, M., Sirocko, F., and Zorita, E.: European summer temperature response to annually dated volcanic eruptions over the past nine centuries, Bull. Volcanol., 75, 1–14, 2013. 2652

Fernández-Donado, L., González-Rouco, J. F., Raible, C. C., Ammann, C. M., Barriopedro, D., García-Bustamante, E., Jungclaus, J. H., Lorenz, S. J., Luterbacher, J., Phipps, S. J., Servonnat, J., Swingedouw, D., Tett, S. F. B., Wagner, S., Yiou, P., and Zorita, E.: Large-scale temperature response to external forcing in simulations and reconstructions of the last millennium, Clim. Past, 9, 393–421, doi:10.5194/cp-9-393-2013, 2013. 2630, 2631

Franke, J., Frank, D., Raible, C. C., Esper, J., and Brönnimann, S.: Spectral biases in tree-ring climate proxies, Nat. Clim. Change, 3, 360–364, 2013. 2652

Fritts, H. C.: Tree Rings and Climate, Academic Press, London, 1976. 2640

Goosse, H., Crespin, E., Dubinkina, S., Loutre, M.-F., Mann, M. E., Renssen, H., Sallaz-Damaz, Y., and Shindell, D.: The role of forcing and internal dynamics in explaining the ‘Medieval Climate Anomaly’, Clim. Dynam., 39, 2847–2866, 2012. 2630

Gunnarson, B. E., Linderholm, H. W., and Moberg, A.: Improving a tree-ring reconstruction from west-central Scandinavia: 900 years of warm-season temperatures, Clim. Dynam., 36, 97–108, 2011. 2673, 2675

Hansen, J., Ruedy, R., Sato, M., and Lo, K.: Global surface temperature change, Rev. Geophys., 48, RG4004, doi:10.1029/2010RG000345, 2010. 2638, 2639

Hasselmann, K.: Stochastic climate models. Part I. Theory, Tellus, 28, 473–485, 1976. 2635

Hegerl, G. C., Crowley, T. J., Hyde, W. T., and Frame, D. J.: Climate sensitivity constrained by temperature reconstructions over the past seven centuries, Nature, 440, 1029–1032, 2006. 2630

Hegerl, G. C., Crowley, T. J., Allen, M., Hyde, W. T., N., P. H., Smerdon, J., and Zorita, E.: Detection of human influence on a new, validated 1500-year temperature reconstruction, J. Climate, 20, 650–666, 2007. 2630, 2651, 2652

Hegerl, G. C., Luterbacher, J., González-Rouco, F., Tett, S. F. B., Crowley, T., and Xoplaki, E.: Influence of human and natural forcing on European seasonal temperatures, Nat. Geosci., 4, 99–103, 2011. 2630, 2651

Hind, A. and Moberg, A.: Past millennial solar forcing magnitude. A statistical hemispheric-scale climate model versus proxy data comparison, Clim. Dynam., 41, 2527–2537, 2013. 2631, 2635, 2637, 2638, 2646, 2650, 2651

Hind, A., Moberg, A., and Sundberg, R.: Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium – Part 2: A pseudo-proxy study addressing the amplitude of solar forcing, Clim. Past, 8, 1355–1365, doi:10.5194/cp-8-1355-2012, 2012. 2630, 2631, 2634, 2635, 2637, 2638, 2650, 2651, 2653, 2659

Hughes, M. K.: Dendrochronology in climatology – the state of the art, Dendrochronologia, 20, 95–116, 2002. 2640

Hughes, M. K., Swetnam, T. W., and Diaz, H. F.: Dendroclimatology, Progress and Prospects, Springer, Dordrecht, Heidelberg, London, New York, 2011. 2640

Jones, P. D., Briffa, K. R., Osborn, T. J., Lough, J. M., van Ommen, T. D., Vinther, B. M., Luterbacher, J., Wahl, E. R., Zwiers, F. W., Mann, M. E., Schmidt, G. A., Ammann, C. M., Buckley, B. M., Cobb, K. M., Esper, J., Goosse, H., Graham, N., Jansen, 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. 2630, 2632, 2640

Jones, P. D., Melvin, T. M., Harpham, C., Grudd, H., and Helama, S.: Cool north European summers and possible links to explosive volcanic eruptions, J. Geophys. Res.-Atmos., 118, 6259–6265, 2013. 2652

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. 2630, 2631, 2635, 2637, 2638, 2650

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. 2638

Landrum, L., Otto-Bliesner, B. L., Wahl, E. R., Conley, A., Lawrence, P. J., Rosenbloom, N., and Teng, H.: Last millennium climate and its variability in CCSM4, J. Climate, 26, 1085–1111, 2012. 2630

Lockwood, M.: Shining a light on solar impacts, Nat. Clim. Change, 1, 98–99, 2011. 2631

Luckman, B. and Wilson, R.: Summer temperatures in the Canadian Rockies during the last millennium: a revised record, Clim. Dynam., 24, 131–144, 2005. 2673

Masson-Delmotte, V., Schulz, M., Abe-Ouchi, A., Beer, J., Ganopolski, A., González Rouco, J. F., Jansen, E., Lambeck, K., Luterbacher, J., Naish, T., Osborn, T., Otto-Bliesner, B., Quinn, T., Ramesh, R., Rojas, M., Shao, X., and Timmermann, A.: Information from Paleoclimate Archives, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, UK and New York, NY, USA, 2013. 2631, 2651, 2652

Matlab: Statistics Toolbox User's Guide R2008a, The MathWorks, Inc., Natick, MA, USA, 2008. 2644

Melvin, T. M. and Briffa, K. R.: A "signal-free" approach to dendroclimatic standardisation, Dendrochronologia, 26, 71–86, 2008. 2640

Melvin, T. M., Grudd, H., and Briffa, K. R.: Potential bias in 'updating' tree-ring chronologies using regional curve standardisation: re-processing 1500 years of Torneträsk density and ring-width data, Holocene, 23, 364–373, 2013. 2673

Morice, C. P., Kennedy, J. J., Rayner, N. A., and Jones, P. D.: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HadCRUT4 data set, J. Geophys. Res.-Atmos., 117, D08101, doi:10.1029/2011JD017187, 2012. 2638, 2639

Osborn, T. J., Briffa, K. R., and Jones, P. D.: Adjusting variance for sample size in tree-ring chronologies and other regional mean timeseries, Dendrochronologia, 15, 89–99, 1997. 2641, 2652

PAGES2k Network: Continental-scale temperature variability during the past two millennia, Nat. Geosci., 6, 339–346, 2013. 2653

Phipps, S. J., McGregor, H. V., Gergis, J., Gallant, A. J. E., Neukom, R., Stevenson, S., Ackerley, D., Brown, J. R., Fischer, M. J., and van Ommen, T. D.: Paleoclimate data–model comparison and the role of climate forcings over the past 1500 years, J. Climate, 26, 6915–6936, 2013. 2635

Schmidt, G. A.: Enhancing the relevance of palaeoclimate model/data comparisons for assessments of future climate change, J. Quaternary Sci., 25, 79–87, 2010. 2630

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. 2630

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 (v1.1), Geosci. Model Dev., 5, 185–191, doi:10.5194/gmd-5-185-2012, 2012. 2631

Schurer, A., Hegerl, G., Mann, M., Tett, S., and Phipps, S.: Separating forced from chaotic climate variability over the past millennium, J. Climate, 26, 6954–6973, 2013. 2652

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. 2652

Smith, T. M., Reynolds, R. W., Peterson, T. C., and Lawrimore, J.: Improvement to NOAA's historical merged land-ocean surface temperature analysis (1880–2006), J. Climate, 21, 2283–2296, 2008. 2638

St. George, S. and Ault, T.: The imprint of climate within Northern Hemisphere trees, Quaternary. Sci. Rev., 89, 1–4, 2014. 2640

Sueyoshi, T., Ohgaito, R., Yamamoto, A., Chikamoto, M. O., Hajima, T., Okajima, H., Yoshimori, M., Abe, M., O'ishi, R., Saito, F., Watanabe, S., Kawamiya, M., and Abe-Ouchi, A.: Set-up of the PMIP3 paleoclimate experiments conducted using an Earth system model, MIROC-ESM, Geosci. Model Dev., 6, 819–836, doi:10.5194/gmd-6-819-2013, 2013. 2630

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. 2630, 2650

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

Vyushin, D. I., Kushner, P. J., and Zwiers, F.: Modeling and understanding persistence of climate variability, J. Geophys. Res. Atmos., 117, D21106, doi:10.1029/2012JD018240, 2012. 2635, 2647, 2650

Wanner, H., Beer, J., Bütikofer, J., Crowley, T., Cubasch, U., Flückiger, J., Goosse, H., Grosjean, M., Joos, F., Kaplan, J., Küttel, M., Müller, S., Prentice, I., Solomina, O., Stocker, T., Tarasov, P., Wagner, M., and Widmann, M.: Mid- to Late Holocene climate change: an overview, Quaternary. Sci. Rev., 27, 1791–1828, 2008. 2630

Widmann, M., Goosse, H., van der Schrier, G., Schnur, R., and Barkmeijer, J.: Using data assimilation to study extratropical Northern Hemisphere climate over the last millennium, Clim. Past, 6, 627–644, doi:10.5194/cp-6-627-2010, 2010. 2630

Wilson, R., Wiles, G., D'Arrigo, R., and Zweck, C.: Cycles and shifts: 1300 years of multi-decadal temperature variability in the Gulf of Alaska, Clim. Dynam., 28, 425–440, 2007. 2641, 2673

Wilson, R. J. S.: Eurasian Regional Composite Chronologies, Research report prepared for Rosanne D'Arrigo and Gordon Jacoby of the Lamont Doherty Earth Observatory (LDEO) Tree-Ring Laboratory, Columbia University, New York, 31 pp., 2004. 2673, 2675

Table 1. Tree-ring temperature reconstructions used in this study, with seasonal representation as determined by the respective investigators. TRW – tree-ring width, MXD – maximum density. IND – individual standardization, RCS – regional curve standardization, SF – signal-free standardization. Short names and start year used in this study are also given.

Name	Abbr.	Proxy	Stand.	Season	Start	Reference
Gulf of Alaska	GOA	TRW	IND	Jan–Sep	1000	Wilson et al. (2007)
Firth River	FIRTH	MXD	RCS + SF	Jul–Aug	1073	Anchukaitis et al. (2013)
Coppermine/Thelon[1]	CT	MXD	IND	May–Aug	1492	D'Arrigo et al. (2009)
Canadian Rockies	CANR	MXD	RCS	May–Aug	1000	Luckman and Wilson (2005)
Torneträsk	TORN	MXD	RCS+SF	May–Aug	1000	Melvin et al. (2013)
Jämtland	JAMT	MXD	RCS	Apr–Sep	1107	Gunnarson et al. (2011)
Tatra	TATRA	TRW	RCS	May–Jun	1040	Büntgen et al. (2013)
Alps	ALPS	MXD	RCS	Jun–Sep	1000	Büntgen et al. (2006)
Pyrenees	PYR	MXD	RCS	May–Sep	1260	Dorado Liñán et al. (2012)
Yamalia Combined	YAMC	MXD + TRW	RCS + SF	Jun–Jul	1000	Briffa et al. (2013)
Avam-Taimyr	AVAMT	TRW	RCS	Jul	1000	Briffa et al. (2008)
Yakutia[2]	YAK	TRW	RCS	Jun–Jul	1342	D'Arrigo et al. (2006)
East Asia[3]	ASIA2K	TRW	other	Jun–Aug	1000	Cook et al. (2013)
Tasmania	TASM	TRW	RCS	Nov–Apr	1000	Cook et al. (2000)
New Zealand	NZ	TRW	RCS	Jan–Mar	1000	Cook et al. (2002, 2006)

[1] Arithmetic average of normalized Coppermine and Thelon data.

[2] Seasonal representation as in Wilson (2004).

[3] Detailed information on standardization is not provided in Cook et al. (2013), but included a partial use of SF.

Table 2. Data sources for the tree-ring records.

Record	Source
GOA	ftp://ftp.ncdc.noaa.gov/pub/data/paleo/treering/reconstructions/gulf_of_alaska/goa2007temp.txt
FIRTH	ftp://ftp.ncdc.noaa.gov/pub/data/paleo/treering/reconstructions/northamerica/usa/alaska/firth2013temperature.txt
CT	ftp://ftp.ncdc.noaa.gov/pub/data/paleo/treering/reconstructions/northamerica/usa/alaska/firth2013temperature.txt
CANR	ftp://ftp.ncdc.noaa.gov/pub/data/paleo/treering/reconstructions/canada/icefields-summer-maxt.txt
TORN	http://www.cru.uea.ac.uk/cru/papers/melvin2012holocene/TornFigs.zip
JAMT	ftp://ftp.ncdc.noaa.gov/pub/data/paleo/treering/reconstructions/europe/sweden/gunnarson2011temp.txt
TATRA	ftp://ftp.ncdc.noaa.gov/pub/data/paleo/treering/reconstructions/europe/tatra2013temp.txt
ALPS	ftp://ftp.ncdc.noaa.gov/pub/data/paleo/treering/reconstructions/europe/buentgen2011europe.txt
PYR	ftp://ftp.ncdc.noaa.gov/pub/data/paleo/pages2k/DatabaseS1-All-proxy-records.xlsx
YAMC	ftp://ftp.ncdc.noaa.gov/pub/data/paleo/treering/reconstructions/asia/russia/yamalia2013temp1000yr.txt
AVAMT	http://www.cru.uea.ac.uk/cru/papers/briffa2008philtrans/Column.prn
YAK	ftp://ftp.ncdc.noaa.gov/pub/data/paleo/treering/reconstructions/n_hem_temp/nhtemp-darrigo2006.txt
ASIA2K	ftp://ftp.ncdc.noaa.gov/pub/data/paleo/pages2k/DatabaseS2-Regional-Temperature-Reconstructions.xlsx
TASM	ftp://ftp.ncdc.noaa.gov/pub/data/paleo/treering/reconstructions/tasmania/tasmania_recon.txt
NZ	ftp://ftp.ncdc.noaa.gov/pub/data/paleo/pages2k/DatabaseS1-All-proxy-records.xlsx

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

Table 3. Selected information for each region: Latitude/longitude boundaries, proxy data calibration period and correlation r with instrumental target, area weight c_j, cluster weight c_j. The area weight is defined as the fraction of the global area. The cluster weight is explained in the text.

Name	lat.	lon.	cal. period	r	area c_j	cluster c_j
GOA	59–63° N	135–161° W	1899–1985	0.55	0.12	1/4
FIRTH	67–71° N	125–149° W	1897–2002	0.63	0.08	1/4
CT	61–71° N	93–121° W	1950–1979	0.75	0.28	1/4
CANR	47–59° N	109–125° W	1895–1994	0.62	0.28	1/4
TORN	63–71° N	11–25° E	1880–2006	0.78	0.11	1/2
JAMT[1]	59–67° N	7–27° E	1880–2007	0.75	0.18	1/2
TATRA	45–53° N	13–29° E	1901–2009	0.42	0.20	1/3
ALPS	45–47° N	3° W–11° E	1911–2003	0.72	0.03	1/3
PYR	39–45° N	7° W–5° E	1900–2005	0.64	0.13	1/3
YAMC	61–75° N	53–81° E	1883–2005	0.79	0.36	1/2
AVAMT	65–75° N	81–107° E	1950–1994	0.66	0.22	1/2
YAK[2]	65–73° N	137–161° E	1951–1980	0.70	0.17	1
ASIA2K	37–55° N	73–143° E	1951–1989	0.59	2.11	1
TASM	39–49° S	127–159° E	1886–1991	0.52	0.56	1/2
NZ	39–47° S	163–177° E	1894–1957	0.45	0.20	1/2

[1] Gunnarson et al. (2011) used 1870–2007, but GISS1200 starts in 1880.

[2] Calibration period as in Wilson (2004).

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

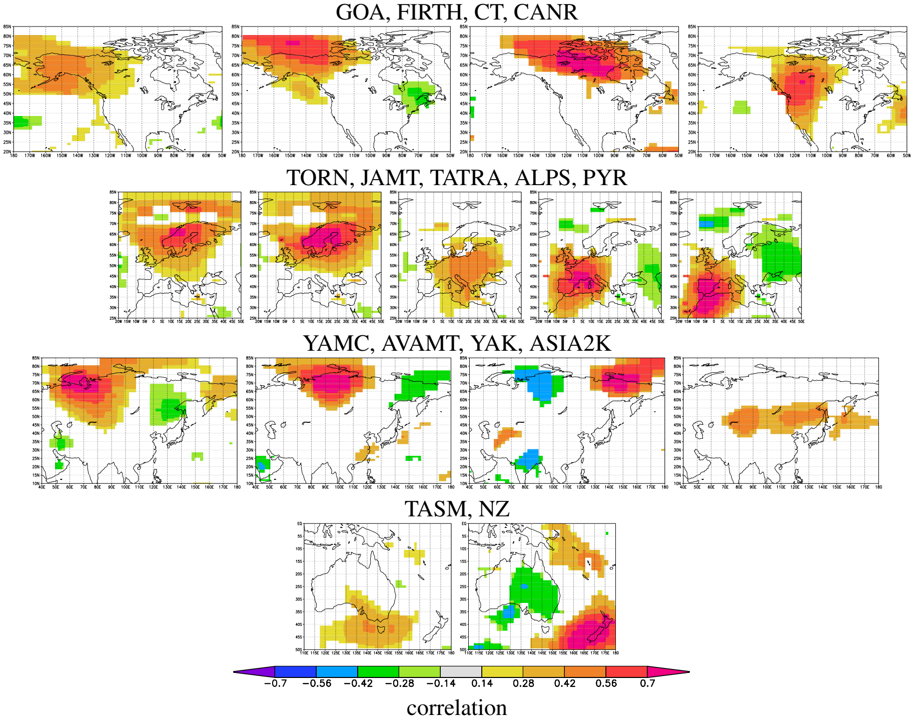

Figure 1. Correlation between each tree-ring chronology and the GISS1200 instrumental temperature field, based on first-differenced data for seasonal averages and time periods as used for calibration by each original investigator (see Tables 1 and 3). Colours are muted where correlations are not significant at the 5 % level. Analysis made on Climate Explorer (http://climexp.knmi.nl).

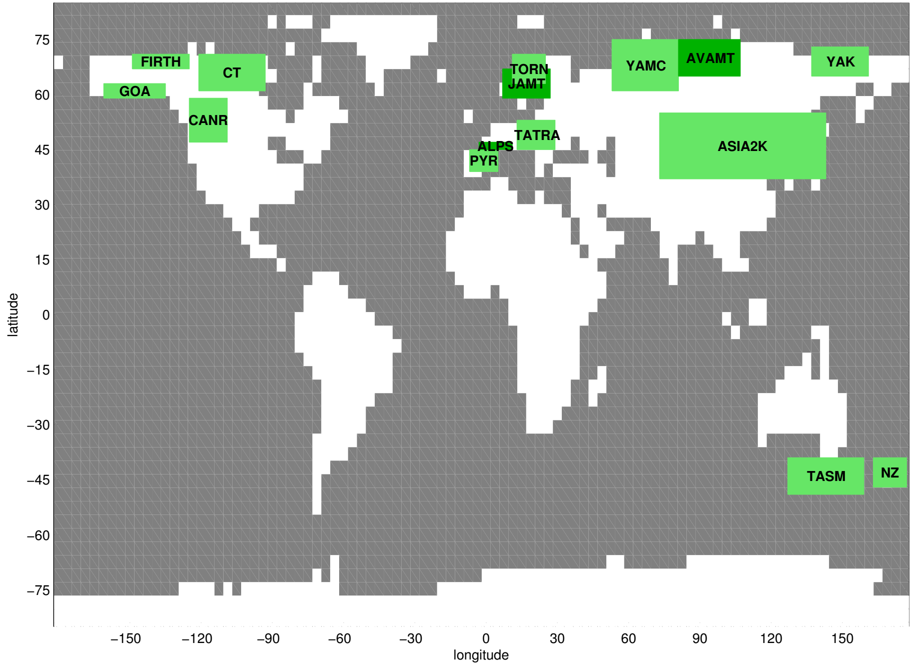

Figure 2. Location of regions that the fifteen tree-ring records represent, plotted on the land/sea mask of the MPI-ESM model. Regions' short names are explained in Table 1.

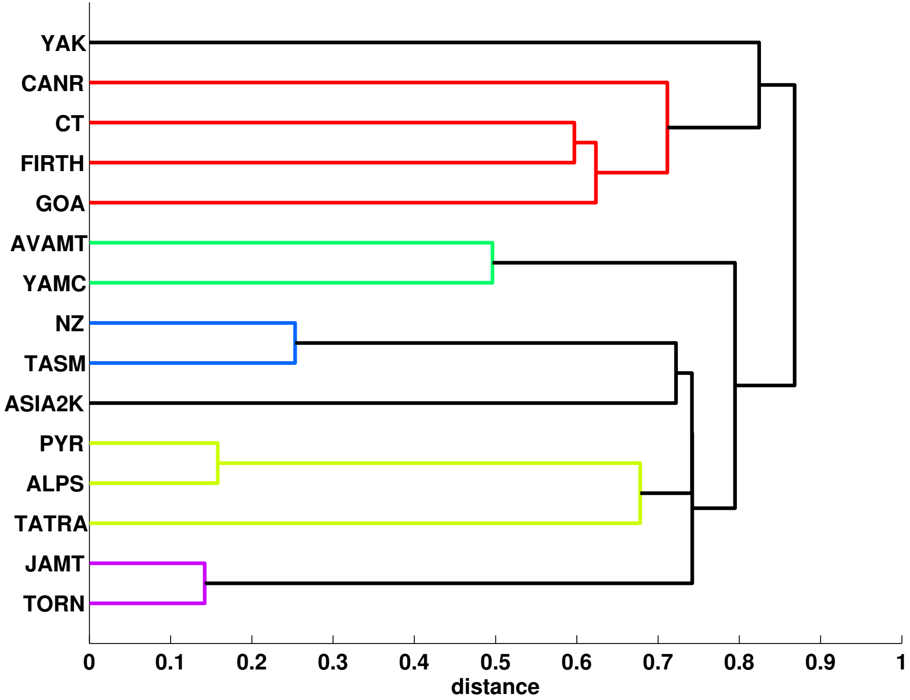

Figure 3. Hierarchical cluster tree based on nearest neighbour linkage with (1 – *r*) as distance metric, where *r* is the sample correlation. Data from the 3000 year long unforced control simulation, for seasons as specified in Table 1, are used for the cluster analysis.

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

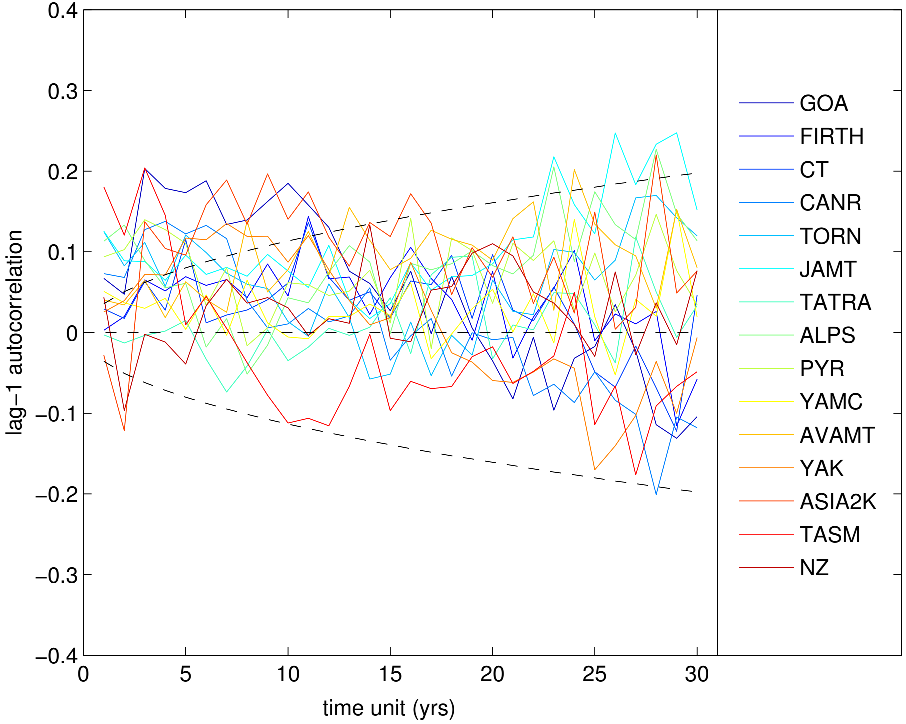

Figure 4. Estimated lag-1 autocorrelation for time units from 1 to 30 years in the 3000 years long unforced control simulation. Two-sided 5 % significance levels for a white noise process are shown with dashed lines. Data for each region, identified by the colour legend to the right, are for the season as specified in Table 1.

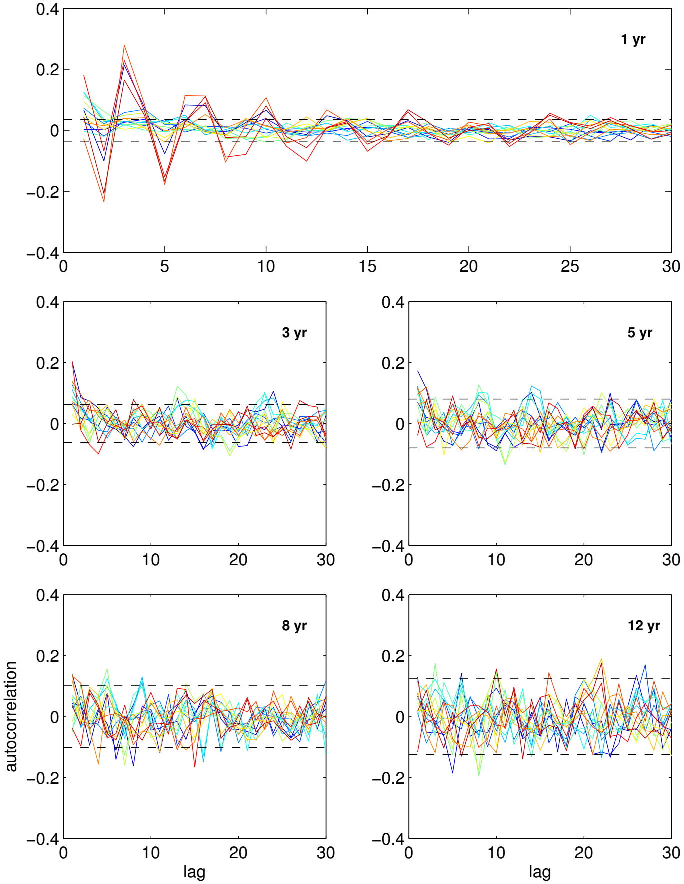

Figure 5. Estimated autocorrelation function for lags up to 30 in the 3000 years long unforced control simulation, for time units of 1, 3, 5, 8 and 12 years. Two-sided 5 % significance levels for a white noise process are shown with dashed lines. Data for each region, identified by the colour legend to the right in Fig. 4, are for the season as specified in Table 1.

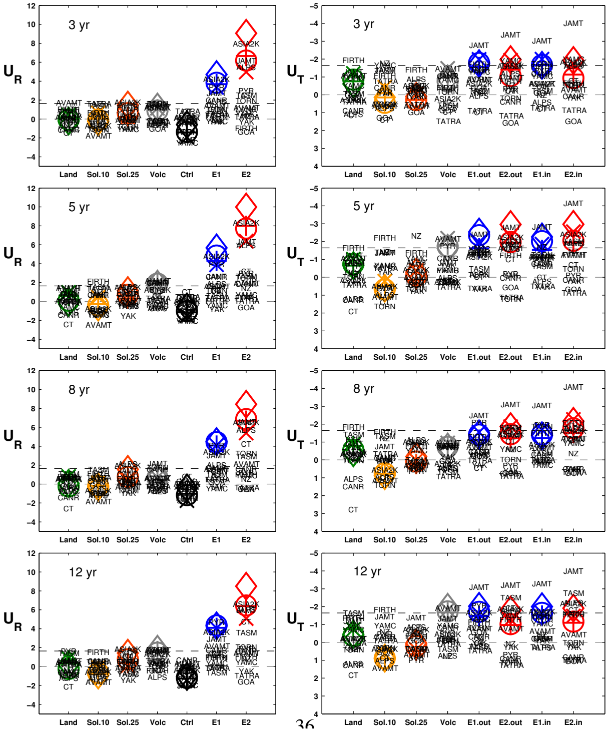

Figure 6. U_R and U_T statistics from comparisons between simulated temperatures and tree-ring-based temperature observations in the period 1000–1849 CE, for time units of 3, 5, 8 and 12 years. Results are shown for single-forcing simulations (land-use, low-amplitude solar, high-amplitude solar, volcanic) and the E1 and E2 multiple-forcing ensembles (with both "outside" and "inside" averaging for U_T). U_R values are also shown for the Ctrl simulation. U_R and U_T values for each region are denoted with site short names. Results where all sites are combined are shown with symbols to distinguish between different c_j weightings (○ equal, ◊ area, + cluster, × equal without non-RCS series). Upper dashed lines show 5 % significance levels. Note the reversed vertical axis in the U_T graphs.

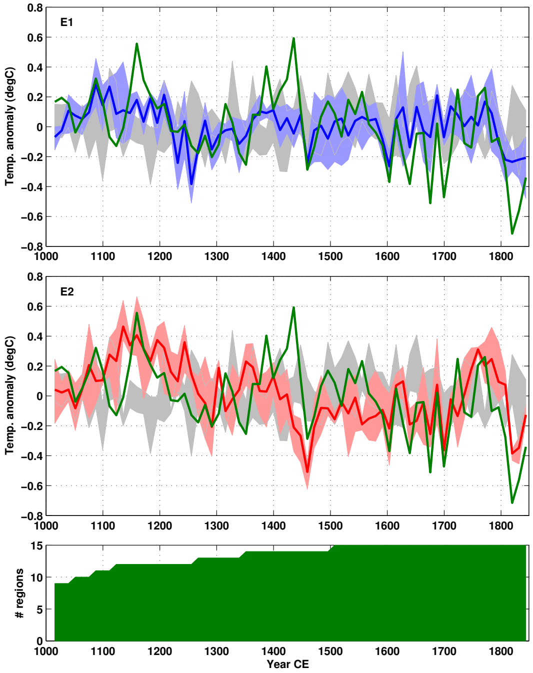

Figure 7. Time-series illustration of data used for the U_R and U_T analysis at the 12 year time unit. Green curves show the arithmetic average of all calibrated proxy series (z series). Blue and red curves show the corresponding values for simulated temperatures (x series), additionally averaged over the E1 (blue) and E2 (red) ensemble members. Light-blue and light-red bands show the range between the highest and lowest regionally averaged simulated temperatures within the E1 and E2 ensembles. Grey bands show the corresponding range for the Ctrl simulation ensemble. Temperatures are shown as anomalies with respect to long-term averages, as used for the U_R and U_T calculations. The bottom graph shows how the number of regions change with time.

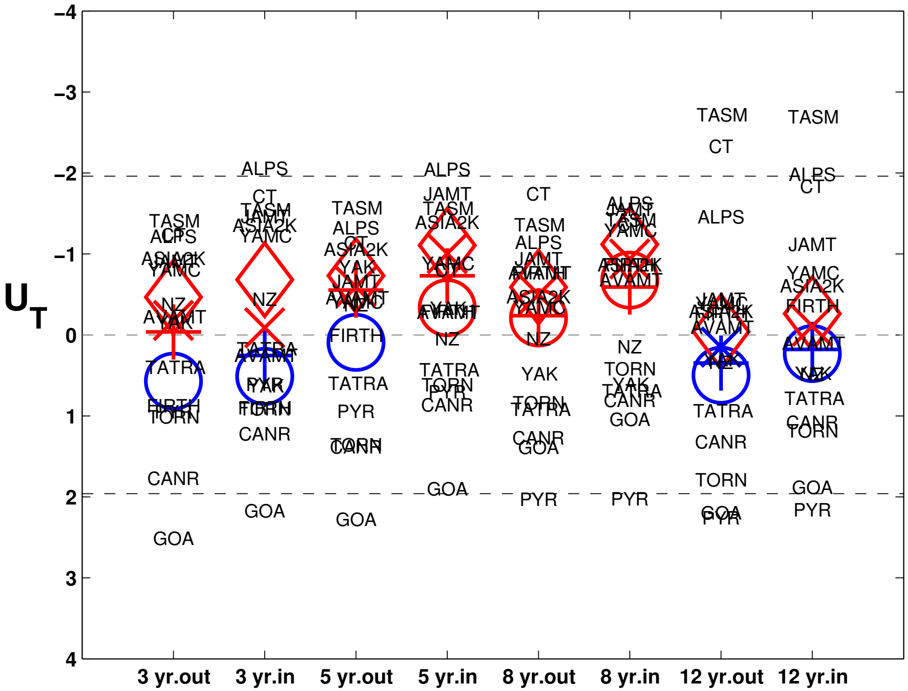

Figure 8. U_T statistics comparing the E2 and E1 multiple-forcing simulation ensembles. Negative values (upwards) indicate where E2 is closer than E1 to the tree-ring based observations. Symbols show results where all regions are combined with different c_j weightings (○ equal, ◊ area, + cluster, × equal without non-RCS series), with colours highlighting where E2 (red) or E1 (blue) is closer to the observations. Results are shown for "outside" and "inside" averaging and four time units. Black dashed lines show 5 % significance levels for testing the null hypothesis that no model is better than the other.

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