Clim. Past Discuss., 8, 263–320, 2012
www.clim-past-discuss.net/8/263/2012/
doi:10.5194/cpd-8-263-2012

Climate of the Past Discussions

Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium

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

[1]Department of Physical Geography and Quaternary Geology, Bert Bolin Centre for Climate Research, Stockholm University, 10691 Stockholm, Sweden
[2]Department of Mathematics, Division of Mathematical Statistics, Stockholm University, 10691 Stockholm, Sweden

Received: 13 December 2011 – Accepted: 29 December 2011 – Published: 12 January 2012

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

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

Abstract

A statistical framework for comparing the output of ensemble simulations from global climate models with networks of climate proxy and instrumental records is developed, focusing on near-surface temperatures for the last millennium. This framework includes the formulation of a joint statistical model for proxy data, instrumental data and simulation data, which is used to optimize a quadratic distance measure for ranking climate model simulations. An essential underlying assumption is that the simulations and the proxy/instrumental series have a shared component of variability that is due to temporal changes in external forcing, such as volcanic aerosol load, solar irradiance changes and greenhouse gas concentrations. Two statistical tests are formulated. Firstly, a preliminary test to establish whether a significant temporal correlation exists between instrumental/proxy and simulation data. Secondly, the distance measure is expressed in the form of a test statistic of whether a forced simulation is closer to the instrumental/proxy series than unforced simulations. The proposed framework allows any number of proxy locations to be used jointly, with different seasons, record lengths and statistical precision. The new methods are applied in a pseudo-proxy experiment. Here, a set of previously published millennial forced model simulations, including both "low" and "high" solar radiative forcing histories together with other common forcings, were used to define "true" target temperatures as well as pseudo-proxy and pseudo-instrumental series. The pseudo-proxies were created to reflect current proxy locations and noise levels, where it was found that the low and high solar full-forcing simulations could be distinguished when the latter were used as targets. When the former were used as targets, a greater number of proxy locations were needed to make this distinction. It was also found that to improve detectability of the low solar simulations, increasing the signal-to-noise ratio was more efficient than increasing the spatial coverage of the proxy network. In the next phase of the work, we will apply these methods to real proxy and instrumental data, with the aim to distinguish which of the two solar forcing histories is most compatible with the observed/reconstructed climate.

CPD
8, 263–320, 2012

1 Introduction

Studies that compare climate reconstructions for the last millennium with climate model simulations have contributed significantly to our understanding of natural and anthropogenic climate change. Based upon results from such investigations, the Intergovernmental Panel on Climate Change concluded in its fourth assessment report, that volcanic and solar forcings have very likely affected NH mean temperature over the past millennium, that external influences explain a substantial fraction of inter-decadal temperature variability in the past; and that the climate response to greenhouse gas increases can be detected in a range of multi-proxy reconstructions during recent decades (Hegerl et al., 2007b). More recently, detection of temperature changes and their attribution to external influences, such as the concentration of stratospheric aerosols and possibly changes in total solar irradiance, have been made at a regional (European) scale for the last five centuries (Hegerl et al., 2011). Moreover, spatial patterns of temperature and precipitation changes, as well as the movement of the intertropical convergence zone, have began to be understood in terms of dynamical responses to natural radiative forcing changes (Mann et al., 2009; Sachs et al., 2009; Graham et al., 2011). A growing size of climate model simulation ensembles for the last millennium (e.g. Jungclaus et al., 2010) and a constantly increasing number of local/regional climate reconstructions from proxy data (Jones et al., 2009) will make it possible to undertake a more systematic evaluation of model simulations against proxy data. However, the growing amount of information also calls for new statistical tools for evaluating the models against the reconstructions. Statistical measures of model performance in terms of mean square errors have long since been used within weather prediction to compare different forecast systems and to track forecast improvements over time (Krishnamurti et al., 1999). These ideas have developed into methods for the detection and attribution of climate change signals using the instrumental record (Allen and Tett, 1999) and palaeoclimate reconstruction data (Hegerl et al., 2007a), as well as techniques for data assimilation of climate proxy data in model simulations

CPD
8, 263–320, 2012

(Goosse et al., 2006; Widmann et al., 2010). Statistical measures of climate model performance can use spatial correlations found in natural climate variability and also combine information from several climate field variables (Mu et al., 2004). However, explicit treatment of the model and observational data error terms in the formulation of performance metrics becomes a great challenge when dealing with climate proxy data because they are typically associated with substantial uncertainties, including mixed seasonal signals and time-scale dependent, temporally unstable climate-proxy relationships. Moreover, the available proxy data are irregularly distributed in space, vary in seasonal representativeness and can reflect different climate variables (Jones et al., 2009). Our aim is to address some of these problems and formulate a statistical framework for evaluation of climate model simulations against a diverse set of climate proxy series. We will assume that evaluation of the models against modern instrumental gridded data sets has already been made and that the models to be tested have been judged to simulate the current climate conditions reasonably well. Hence we focus on problems connected with how to use climate proxy data for model evaluation back into the pre-instrumental period. We demand that the proxies have sufficiently high temporal resolution and dating precision to allow direct calibration against instrumental climate time series. In practice, this requirement excludes many types of proxy data and also time periods beyond the last millennium. Tree-ring data and historical documentary proxies are annually resolved and have exact dating, which make them suitable. Some proxies with lower resolution, but still with a great deal of precision in their dating, may also be considered provided that their sampling resolution is high enough to allow meaningful calibration against overlapping instrumental series.

We start by formulating a statistical model, where we have near-surface temperatures in mind, from which a climate model evaluation framework is developed. Note that other climate variables, such as precipitation or a drought index, are probably more difficult to model and may require substantial modification of the theory presented below. To investigate the performance of our framework, we undertake a pseudo-proxy experiment where we study the possibility to distinguish between climate model simulations

CPD
8, 263–320, 2012

that use different past climate forcings. The purpose with this exercise is to learn, under controlled conditions, about the possibility to identify simulations that show a statistically significant fit to real climate proxy data and, when several simulations achieve this, to rank them according to their goodness of fit. This knowledge will help understanding of how to interpret results obtained when the simulations are compared with real proxy data.

2 Two statistical models

We assume the climate characteristic of interest, to be called τ, is a temperature time series representing a particular region during some time period, divided into a number of time units yielding a sequence of values τ_i, $i = 1, ..., n$, where the subscript i represents time. Typically this region consists of a single model grid-box, but averages over several grid-boxes can also be considered. The time unit can be single years or equally say, averages over a ten-year or thirty-year period. To begin with, we only consider temperatures for a single region and a particular season, but later (in Sect. 7) we will investigate how to combine data from different regions and seasons. Let

- x = a simulated temperature value for the region and time period of interest, generated by a climate model.
- τ = a true temperature, corresponding to x; a spatial and temporal average over the region and the time unit. The true temperature is an unobserved (or latent) variable, except in those cases where we set $\tau = y$, see below.
- y = a measured temperature, intended to represent τ, being also some average over space and time, and available for some period of time. This measured value y can differ from τ because of measurement errors, but also because y and τ are somewhat different spatial and temporal averages (typically, y is an average taken over a finite set of irregularly spread observing stations and for a set of possibly

CPD
8, 263–320, 2012

time-varying observation hours). Sometimes we will assume that this observed temperature well enough approximates the true temperature τ, so we can neglect measurement type errors in these observations. However, often in practice we expect some non-negligible errors to exist.

– z = a proxy for the true temperature, τ. When observed temperatures y are not available, proxy measurements will be used. Here we ignore all practical problems connected with how to construct temperature proxy series from raw proxy data (e.g. tree-ring width or density measurements). Hence, we think of a proxy series as a final product for use in climate reconstruction (e.g. a tree-ring chronology), constructed in the best possible way.

The following statistical model explicitly allows climatic forcing effects jointly in the climate model simulations and in the actual temperature. This is crucial, since inclusion of temporally varying external forcings in the climate model simulation is the only reason to expect any temporal correlation between simulations and actual temperature. The forcing effects can, for example, be the temperature response to radiative forcing from stratospheric aerosols ejected from large volcanic eruptions or the response to variations in solar radiation. Note that any type of forcing imposed on the climate model is not a true reflection of reality because the forcing history is incompletely known regarding its temporal evolution, its amplitude and its spatial distribution pattern. Moreover, it is typically only crudely represented in the simulations. Its effect on temperature need not be the same in reality as in the model because these worlds may have different sensitivities to the forcing and possibly also different spatial response patterns. For simplicity we will assume that the relationship between the true and simulated forcing effect is (approximately) proportional, when measured as deviations from the mean values of τ and x, but with an unknown proportionality constant.

CPD
8, 263–320, 2012

Statistical Model 1 Climate model simulation sequence $\{x_i\}$, true climate sequence $\{\tau_i\}$, instrumental measurement sequence $\{y_i\}$, and proxy sequence $\{z_i\}$ are mutually related through the following model, explained below:

- $x_i = \mu_x + \alpha\xi_i + \delta_i$
- $\tau_i = \mu_\tau + \xi_i + \eta_i$
- $y_i = \tau_i + \theta_i$
- $z_i = \mu_z + \beta(\tau_i - \mu_\tau) + \epsilon_i$

Here, Greek letters are used for latent variables, random variables, unobserved errors and unknown coefficients, to indicate their unobservability. In contrast, x, y and z are observed or measured. Terms μ_x, μ_τ and μ_z are the mean values over time, around which x, τ (and y), and z vary. Quantities ξ, δ, η, θ, and ϵ are regarded as random variables, with mean values zero and variances σ^2_ξ, σ^2_δ, etc., whereas α and β are unknown coefficients:

- ξ denotes the true effect of a specific type of forcing that has influenced the true temperature τ. Since both the causes behind the forcings and the actual effects are uncontrolled, we regard this variation as random. The forcing can be either of a single-type (e.g. only volcanic forcing) or a combination of several forcings (e.g. volcanic and solar forcing). Note that ξ is *not* the forcing itself, but rather its temperature response.
- $\alpha\xi$ represents the unknown variability in x that is due to the forcing imposed on the simulation. For simplicity we have assumed an (approximately) proportional relationship to the true effect on τ. A correct representation of the forcing effect in the climate model corresponds to $\alpha = 1$.
- η denotes the (residual) variation in true temperature that is not due to the particular forcing under consideration. This is supposedly uncorrelated with ξ.

CPD
8, 263–320, 2012

– θ denotes the measurement error in the temperature variable y, making y differ from the true temperature τ.

– δ represents internal noise variability in x, i.e. variability not due to the forcing included in the simulation. It will also incorporate nonlinear forcing effects that are orthogonal (uncorrelated) to ξ.

– $\beta(\tau_i - \mu_\tau)$ is the regression of the proxy z on the true temperature τ. The observed proxy value z will be correlated with the measured temperature y, due to the τ they have in common, and we will use that correlation to calibrate the proxy variable.

– ϵ represents the residual variation in z, uncorrelated with y.

It is judged reasonable that all random variables ξ, η, θ, δ and ϵ should be mutually uncorrelated, and this is also assumed below. Under this assumption, a positive correlation between x and τ (or y or z) implies that they share the term ξ. In other words; the effect of the forcing in x corresponds with that in the true temperature τ.

The ξ and η (i.e. the components of τ) sequences will certainly show autocorrelation on various time scales and our theory allows this. Sometimes it may be necessary to consider the more complicated case considering the individual effects of separate multiple forcings, represented by a vector $\boldsymbol{\xi}$ instead of a scalar ξ. Although climate model simulations driven by multiple forcings are used in the experimental Sect. 9, the theoretical aspects of the case with separate multiple forcings will be investigated further in a future analysis. The sequences δ, θ and ϵ will be assumed to be temporally uncorrelated, i.e. white noise, where a specification is needed. This could be a limitation in the theory, because the real processes they represent may all show autocorrelation. In a future refinement of the model, they could be taken to be AR(1) or as having some other time series structure. In the pseudo-proxy experiment part of the present study, however, we will specify both θ and ϵ to be white noise. Concerning the simulated unforced temperature variability δ (in x), we hypothesize that autocorrelation only occurs at rather short timescales and is negligible at longer time scales. We investigate

CPD
8, 263–320, 2012

this problem in Sect. 9 in order to find a time unit length which will make our statistical model valid.

As a reference, we can also consider an unforced model in which there is no external forcing:

Statistical Model 2 The model for data under *unforced* climate model simulations can be written

- $x_i = \mu_x + \delta_i$
- $\tau_i = \mu_\tau + \eta_i$
- $y_i = \tau_i + \theta_i$
- $z_i = \mu_z + \beta(\tau_i - \mu_\tau) + \epsilon_i$

where δ_i, θ_i and ϵ_i (but not η_i) are regarded as white noise.

We will need the unforced model particularly in Sect. 6, where it will have an important role in testing for significance. Note that the forced component of τ (i.e. ξ in Statistical Model 1) is included here in η. A relevant question in this context is whether σ_δ^2 in Model 1 depends on α or not, in particular if it has the same value in Models 1 and 2. We will not deal with this problem here but, when necessary, simply assume that they are the same. Also, we will not consider here, but leave for future analysis, the more complicated problem of multiple forcings, separately controlled in the climate model but with joint effects (where $\boldsymbol{\xi}$ is a vector rather than a scalar). A somewhat related approach to the problem of comparing climate models with the same types of forcings, but to different degrees, would be to try estimating α. Again, this will be a topic for future study.

One can only expect a correlation between a forced simulation and the actual temperature if the forced simulation is able to explain some of the variability in the real temperatures. Thus, in practice, if we want to test several forced simulations of different types and if we want to rank them according to how well they are able to explain the observed temperatures, it is natural to first test whether a forced simulation can

CPD
8, 263–320, 2012

explain any of the observed temperature variations. Only forcings that are found able to explain some of the observed temperature variations, as indicated by a significance test, are worthwhile studying for determination of the optimal forcing magnitude and for use in calculations of a distance measure. Although a correlation test should therefore be computed before any distance measure is calculated, we start the description of our statistical framework by developing a distance-based performance metric (in Sects. 3–7) before we formulate a correlation-based test (in Sect. 8).

3 The distance measure, $D^2(x, z)$

The problem is to identify, among several forced climate model simulations, a simulation that is able to predict the actual temperature better than the others – and in particular better than unforced model simulations. For comparison of different forced simulations, to find out whose x-sequence of temperatures is best at capturing the real variation in temperatures (τ), we need a criterion. Performance metrics for climate model simulations are typically expressed as some kind of squared difference measure (Mu et al., 2004; Goosse et al., 2005, 2006), and we choose a criterion of this kind.

We postpone the problem with proxy data and assume first that we have the true temperatures τ available. We define the simple distance measure

$$D^2(x, \tau) = \frac{1}{n} \sum_{1}^{n} (x_i - \tau_i)^2 .$$

A statistical motivation for this criterion is obtained by considering D^2 as a mean squared error of prediction (MSEP). The better the climate model represents the forcing effects that underlie the true temperature, the smaller the expected distance between simulations and true temperatures. However, any systematic bias in x will also contribute to D^2. If one has good reason to assume that systematic biases can be neglected for a particular study, then this can be achieved by subtracting the mean

CPD
8, 263–320, 2012

values of x and τ over a common time period. Doing so, however, obviously makes the criterion unsuitable for evaluating systematic model biases; rather it then solely focuses on comparing the temporal evolution of climate model simulations with the true temperature evolution.

Since the true τ_i is not available, we have to replace it by the measured y_i for the period when y is observed and else by a suitable proxy z_i. For notational convenience, we suppress y and write $D^2(x, z)$, where z_i is assumed to be replaced by y_i when y_i is available:

$$D^2(x, z) = \frac{1}{n} \sum_{1}^{n} (x_i - z_i)^2 .$$

Leaving aside how z should be chosen for the moment; it is enough that z satisfies the Statistical Model 1. There is motivation to modify D^2 by giving different weights to different terms of $D^2(x, z)$, depending on how good the available data are. However, this discussion will be postponed to Sect. 5. We will first (in Sect. 4) compare $D^2(x, z)$ with the ideal $D^2(x, \tau)$. We do not want $D^2(x, z)$ to yield a systematically different ranking of a set of different x than that given by $D^2(x, \tau)$ and we will see under what circumstances it does not. The criterion for this can be expressed as a calibration procedure for calibration of the proxy series, which tells us how z should be calibrated for use in $D^2(x, z)$. Later, we will discuss the statistical significance and precision of $D^2(x, z)$ (Sect. 6) and how to combine information from several regions or seasons into a unified model performance metric for each model simulation (Sect. 7).

4 How to use instrumental and proxy data in $D^2(x, z)$ to avoid biased ranking

We assume here that we want to rank different climate model simulations according to their ideal distance measure $D^2(x, \tau)$. However, we only have the surrogate measure $D^2(x, z)$, using the observed temperature variable y (when available) or a proxy measurement z instead of the true temperature τ, and we do not want this to change the

CPD
8, 263–320, 2012

ranking in any systematic way. We first conclude that replacement of τ by y does not introduce any ranking bias. This is seen from the relation

$$(x - y)^2 - (x - \tau)^2 = 2(x - \tau)\theta + \theta^2.$$

Averaging over the noise term θ, we obtain zero for the first term and a constant $\mathrm{Var}(\theta) = E(\theta^2)$ for the second term on the right-hand side. This means that the noise term θ of y does not introduce any ranking bias.

The proxy data z must be calibrated. We assume that we have available a period of both proxy and temperature measurements. If possible, we want the calibration of z to be such that, whatever ξ is, it does not introduce any systematic errors to the ideal ranking. The criterion to achieve this is that the expected value (given $\{\xi_i\}$ and $\{x_i\}$) of the difference $D^2(x, z) - D^2(x, \tau)$ should be free from x.

The general term of the expected D^2 difference is

$$E\left\{(x - z)^2 - (x - \tau)^2\right\} = \{E(x - z)\}^2 - \{E(x - \tau)\}^2 + \mathrm{Var}(x - z) - \mathrm{Var}(x - \tau). \quad (1)$$

The variance part of Eq. (1) can be written

$$\mathrm{Var}(x - z) - \mathrm{Var}(x - \tau) = \mathrm{Var}\{(x - \tau) - (z - \tau)\} - \mathrm{Var}(x - \tau) = \mathrm{Var}(z - \tau),$$

because $x - \tau$ and $z - \tau$ are uncorrelated. Hence, the variance part of Eq. (1) is a constant, free from ξ and x, and thus not contributing to any bias.

The first part of Eq. (1) can be written

$$\{E(x - z)\}^2 - \{E(x - \tau)\}^2 = (x - \mu_z - \beta\xi)^2 - (x - \mu_\tau - \xi)^2.$$

The demand on z for this to be free of x is clearly that $\mu_z = \mu_\tau (= \mu_y)$ and that $\beta = 1$, making the two terms cancel. That is, z should first have the same mean value as y. This is naturally achieved in calibration by adding a constant to z so its average value $\bar{z}$ over the calibration period satisfies $\bar{z} = \bar{y}$. Additionally z should have a regression on the latent variable τ with regression coefficient $\beta = 1$. This implies that, given a provisional proxy z_0 (i.e. an uncalibrated proxy) with regression coefficient β_0 on τ, z_0

CPD
8, 263–320, 2012

should be rescaled by β_0, to form a new z, $z = \mu_y + (z_0 - \mu_{z_0})/\beta_0$. In the case when the error in y is negligible, so $y = \tau$, this corresponds precisely to the so called classical calibration procedure (Brown, 1993), when z_0 is regressed on y and this relationship is inverted to yield a predictor/estimator for y.

Next, allow the error in y to be nonnegligible. Then we have a statistical relationship between z_0 and y of the structural relationship type (an errors-in-variables model). Provided that we can estimate or otherwise judge the size of the error variance in y, i.e. σ_θ^2, then we can obtain an approximately unbiased estimator of β_0 by

$$\widehat{\beta}_0 = \frac{s_{yz_0}}{s_y^2 - \widehat{\sigma}_\theta^2}, \tag{2}$$

where s_{yz_0} and s_y^2 are the empirical covariance and variance, respectively. This is the quantity by which to normalize z_0 to obtain the desired z sequence; $z = \mu_y + (z_0 - \mu_{z_0})/\widehat{\beta}_0$. Setting $\widehat{\sigma}_\theta^2 = 0$ brings us back to the previous situation.

Conclusion. To avoid systematic ranking error in the squared distance $D^2(x, z)$ relative to the ideal $D^2(x, \tau)$, the proxy z should be mean adjusted and normalized such that the estimated regression coefficient of z on τ is 1. This corresponds to use of the so called classical calibration procedure for calibrating z against y, if errors in y are negligible, modified to allow errors in y according to Eq. (2), where this is deemed necessary.

Note that in comparison with the observed temperature y, the amplitude of variation in the proxy, Var(z), is exaggerated after classical calibration or when using Eq. (2). The reason is that it should retain the full amplitude of the true temperature signal and that the proxy noise variance is superimposed on the temperature signal variance.

If more than one proxy series is available for the region and season of interest, they should be combined to a single z_0 sequence in order to increase statistical precision and thus yield the smallest possible randomness in $D^2(x, z)$. In theory, this is achieved by multiple regression of y on the set of available proxy series to obtain z_0. In practice, however, there are several reasons (e.g. collinearity among the proxies, or that the

CPD
8, 263–320, 2012

relationships obtained in the calibration period may not hold outside this period) why another way to combine the proxies may be preferred. We will not deal with this practical problem here, but merely conclude that, whatever method chosen, the goal should be to optimize the correlation between z_0 and τ. The preferred z_0 is then rescaled using classical calibration or Eq. (2). In cases when the preliminary proxy series z_0 is known to have different precision in different pre-instrumental time periods, a unique calibration is needed for each such period. Note that this will lead to different variances in the different parts of the final calibrated z series (Sect. 5).

In practice, it is necessary to decide a time unit to use for the calibration. For annually resolved proxy data, the calibration will have its highest precision if calibration is made using the full annual resolution. However, if the model evaluation is made for a lower resolution (e.g. ten or thirty year means) and if there is reason to assume that the proxy/temperature regression relationship is time-scale dependent, then it may be better to use a lower resolution for the calibration but this will of course decrease the statistical precision. The instrumental noise variance to be used in Eq. (2) can be difficult to estimate in practice, but see Moberg and Brattström (2011, Sect. 6.1) for a discussion on a possible procedure.

5 Weighting in $D^2(x, z)$

Direct temperature measurements y and proxies z have mutually different precision. Moreover, the precision (particularly in z) can vary with time due to the quality and quantity of raw data. This motivates giving different weights to different terms (time points) in $D^2(x, z)$. In order to understand how we should introduce weighting in D^2, we first reconsider Statistical Model 1, assuming both $\alpha = 1$ (correct forcing amplitude ξ) and $\beta = 1$ (calibrated z), so that the forcing effect ξ vanishes from $x - z$ and $x - y$. We also assume $\mu_x = \mu_y = \mu_z$, so there is no bias in $x - y$ or $x - z$.

CPD
8, 263–320, 2012

If the climate model is perfect in this sense, and if we first assume a Gaussian distribution with constant variance for the variability of $x - z$, the resulting Gaussian probability density for the observed series $x - z$ is proportional to

$$e^{-\frac{n}{2} D^2(x,\, z)/\left(\sigma_\delta^2 + \sigma_\eta^2 + \sigma_\epsilon^2\right)}, \tag{3}$$

where σ_δ^2, σ_η^2 and σ_ϵ^2 are the variances of the components of the statistical model. When y_i is available and replaces z_i, σ_ϵ^2 should be replaced by σ_θ^2, but for simplicity of notation we leave that alternative aside for the moment. If there is a bias in x and/or a true forcing effect that does not have a linearly correct representation in the climate model (i.e. $\alpha \neq 1$, including the case $\alpha = 0$), its D^2-value will tend to be higher and the probability (Eq. 3) to observe this vector $x - z$ will tend to be exponentially smaller.

The denominator $\sigma_\delta^2 + \sigma_\eta^2 + \sigma_\epsilon^2$ in the exponent of Eq. (3) is a constant. However, when the variances in this denominator vary with i, in particular the proxy noise term $\sigma_\epsilon^2(i)$, the interpretation as a probability tells us how different terms should be (ideally) weighted in D^2, forming a weighted version D_w^2:

$$D_w^2(x,\, z) \;=\; \frac{1}{n} \sum_1^n w_i (x_i \,-\, z_i)^2 \;=\; \frac{1}{n} \sum_1^n \frac{(x_i \,-\, z_i)^2}{\sigma_\delta^2 \,+\, \sigma_\eta^2 \,+\, \sigma_\epsilon^2(i)}.$$

An alternative formulation is to introduce the constant factor $\sigma_\delta^2 + \sigma_\eta^2$, corresponding to use of the density for $x - \tau$ instead of $x - z$ in the numerator of the exponent of Eq. (3). We will use that version as our definition for w_i:

$$w_i \;=\; \frac{\sigma_\delta^2 \,+\, \sigma_\eta^2}{\sigma_\delta^2 \,+\, \sigma_\eta^2 \,+\, \sigma_\epsilon^2(i)}. \tag{4}$$

For times i when a precise y is available (i.e. with $\sigma_\epsilon^2 = \sigma_\theta^2 = 0$), the normalized weight (Eq. 4) equals 1, but $w_i < 1$ when a noisy proxy z is used.

CPD
8, 263–320, 2012

The weight factor introduced in Eq. (4) is an ideal weight, for which we can at best give an estimate. Thus, we must obtain estimates for each of the three components σ_δ^2, σ_η^2 and $\sigma_\epsilon^2(i)$. We assume that the first two components are constant over time, but we allow $\sigma_\epsilon^2(i)$ to vary over time to make it possible to assign different weights at time points (intervals) i when a z with different precision is used.

To estimate σ_δ^2 we propose to use the sample variance s_δ^2, pooled from simulations of an unforced model (control simulation). The more simulations available, the better the estimate will be. The main reason to avoid using forced models here is that their simulations contain an additional source of variation, contributing to the sample variance of the x series. A second reason is that the weights should not differ between the climate models used. The variance σ_η^2 is arguably more difficult to estimate. It represents the unforced real temperature variance, which cannot be estimated directly from instrumental observations (y) because they will always include some forced variance. In particular, the anthropogenic greenhouse gas forcing is likely to be represented as a trend-like component in y which acts to increase the estimated variance of y. Therefore we propose to detrend the observed y before using it to estimate σ_η^2. Fortunately, σ_η^2 (as well as σ_δ^2) occurs in both the numerator and denominator of Eq. (4), so reasonably small errors in its estimate have little influence on the ratio.

Next, we need an estimate of the (possibly) time-varying $\sigma_\epsilon^2(i)$. Although this quantity is needed for time points i outside the calibration period, we estimate it by using information from the calibration period when both y and z are available. Assume first that $y = \tau$, i.e. $\sigma_\theta = 0$. We can use the calibration period to estimate the correlation $\rho(y, z)$. The model formula $z = y + \epsilon$ implies $\rho^2 = \mathrm{Var}(y)/\mathrm{Var}(z)$, from which we obtain the relationship $\sigma_\epsilon^2 = \mathrm{Var}(y)(1-\rho^2)/\rho^2$ (knowing that the regression coefficient of z on y is 1).

Note that this estimate of σ_ϵ^2 is determined by the empirical correlation between the proxy and the instrumental data and therefore by the estimated statistical precision of the proxy. In cases when the proxy series z_i is known to have different precision in different time periods (and hence different calibrations have been made), a unique

CPD
8, 263–320, 2012

weight should be used for each such period, where each weight should be determined by using the corresponding calibration ρ^2. In this way we can allow $\sigma_\varepsilon^2(i)$ in Eq. (4) to vary with time.

Let s_y^2 be the empirical variance of (detrended) y and follow the procedure described above. This yields the weights formula

$$w_i = \frac{s_\delta^2 + s_y^2}{s_\delta^2 + s_y^2/\rho^2(y, z)} \tag{5}$$

for i in the proxy period. Note that for $\rho^2 = 1$ the formula yields $w_i = 1$, as it should do when we use $y = \tau$. As ρ^2 approaches zero, so does w. The higher the ratio s_δ^2/s_y^2, the slower the approach to zero.

Let us now allow noise in y, with noise variance σ_θ^2. If the ratio $q = \sigma_\theta^2/s_y^2 > 0$ is known, the weighting formula for the period when only instrumental data y is used becomes:

$$w_i = \frac{s_\delta^2 + s_y^2(1 - q)}{s_\delta^2 + s_y^2}. \tag{6}$$

In this case the weight is somewhat smaller than 1, depending on the size of q.

For the period when the proxy z is used, the weighting formula becomes:

$$w_i = \frac{s_\delta^2 + s_y^2(1 - q)}{s_\delta^2 + s_y^2(1 - q)^2/\rho^2(y, z)}. \tag{7}$$

A minor drawback of Eq. (7) is that it might generate weights $w_i > 1$. This occurs when the estimated $\rho^2 > 1 - q$ (which is not possible for the *true* values). For that reason w_i could be redefined by using Eq. (5) if this happens. Alternatively the estimation procedures should be checked.

CPD
8, 263–320, 2012

6 Statistical significance and statistical precision of $D^2(x, z)$

Once a D^2 value has been calculated for a forced climate model simulation, for a region and season corresponding to a true temperature series τ, it is relevant to first ask whether this D^2 is better (smaller) than a corresponding D^2 value for an unforced model. To make it possible to answer that question we construct a statistical test of the null hypothesis that the forced model is not better than an unforced model:

H_0: *The climate model under consideration is equivalent with the unforced model.*

Since the unforced model (control simulation) is important here, we recall Statistical Model 2 from Sect. 2. The unforced model is assumed to have been run a number of times, and for each such 'replicate' run (differing in initial conditions, and hence also in the actual trajectories of simulated climate variables) we calculate a D^2 value. Let K denote this number of simulations, and let k denote the number of simulations with a forced model (also differing in initial conditions) where all simulations share the same forcing history. Before we calculate the difference in D^2 between forced and unforced simulations, we average D^2 over all replicates in each of the two terms, respectively. This procedure yields the test statistic

$$T(x_\mathrm{f}, x_\mathrm{u}, z) = \overline{D^2_w}(x_\mathrm{f}, z) - \overline{D^2_w}(x_\mathrm{u}, z) \quad (8)$$

where x_f and x_u represent data from the forced and unforced models, respectively. An alternative averaging procedure would be to take averages over the x series inside each D^2, i.e. to use the average time series $\overline{x_\mathrm{f}}$ and $\overline{x_\mathrm{u}}$ and compute the difference $T(\overline{x_\mathrm{f}}, \overline{x_\mathrm{u}}, z) = D^2_w(\overline{x_\mathrm{f}}, z) - D^2_w(\overline{x_\mathrm{u}}, z)$. This alternative procedure would be even more efficient, but is not used here because it would also introduce a bias in the comparison. However, we provide details necessary to use this alternative in Appendix A.

We show below that an approximate distribution under H_0 for the test statistic in Eq. (8) can be obtained with the help of an analytical formula for its standard error.

CPD
8, 263–320, 2012

In doing this, we will regard the z series as fixed and given. It means that we do not need any distributional asumptions about the z series. This is possible because z is common to both terms of Eq. (8).

Since we are more interested in variation than in mean values, we assume that all x_u and x_f series are mean value adjusted to a common value, that will be denoted μ_x, the test statistic value can be rewritten as

$$T(x_f,\, x_u,\, z) \;=\; \overline{w(x_f \,-\, \mu_x)^2} \;-\; \overline{w(x_u \,-\, \mu_x)^2} \;-\; 2\,\overline{w\,(\overline{x}_f \,-\, \overline{x}_u)(z \,-\, \mu_x)}, \qquad (9)$$

where the overlines in the first two terms represent averaging over both replicates and time index i. Here the factor $(z - \mu_x)$ has the role of a weight factor, multiplying with w. It is natural to adjust the x_u and x_f series additively so that the z series also has the same mean value, $\overline{z} = \mu_x$. Then we write $z - \overline{z}$ in the last term.

The distribution for T is presumably well approximated by a normal distribution, since all terms of the representation Eq. (9) are sums of a large number of terms (referring to the central limit theorem of probability). Under H_0, the expected value of $T(x_f,\, x_u,\, z)$ is zero, since the forced climate model is equivalent with the unforced model. Assuming normality not only of T but already of x_f and x_u, the variance of T can be expressed as

$$\mathrm{Var}(T(x_f,\, x_u,\, z)) \;=\; \frac{1}{n^2}\left(\frac{1}{k} \,+\, \frac{1}{K}\right)\left\{2\,\sigma_\delta^4 \sum_1^n w_i^2 \,+\, 4\,\sigma_\delta^2 \sum_1^n w_i^2 (z_i \,-\, \mu_x)^2\right\}. \qquad (10)$$

An approximately $N(0,\, 1)$-distributed test statistic is obtained by normalizing the T-value in question by its standard error, that is by the square-root of Eq. (10) after insertion of the average $\overline{z}$ for μ_x and of the estimate s_δ^2 for σ_δ^2. It is of some importance to make sure that the estimate s_δ^2 is not too imprecise. As in Sect. 5, we propose to obtain this estimate by calculating the sample variance from all available control simulations.

The test should reject H_0 if the resulting value is too far to the left, e.g. to the left of -1.65 at the 5 % significance level. It should be kept in mind that if many mutually independent climate models are tested against the unforced model, but none of them

CPD
8, 263–320, 2012

has an (appreciable) correlation with the real data y and z, we must nevertheless expect 5 % false positives from tests at the 5 % level, and analogously for the 1 % level, etc. Thus, it is not enough to find one or a few models showing statistical significance, but the whole sequence of model tests must be considered. As a final comment, we note that an alternative way to perform the significance test would be to use a simulated/randomized resampling procedure to empirically determine the distribution of T instead of using the analytical variance formula in Eq. (10), however this is not discussed further here (see Appendix B for details).

7 Combination of data from different seasons and regions

Evaluation of palaeo-simulations from climate models should preferably be made using proxy records from as many regions as possible. Data from different regions and/or seasons could then be combined in a single test, but it is not obvious how this should be done. Proxy records from different regions may represent different seasons and may also be of different lengths. In this section, we define a unified performance metric for each model, based on a normalized sum of test statistics T for all regions/seasons with available proxy data.

This sum of test statistics can be a simple or a weighted one. Weighting could be implemented if we want a balanced spatial average but the regions are of different size or have a different density of proxy values, or if we want a balanced annual average for a region with quite different numbers of summer and winter values. Different *quality* of the proxies does *not* necessitate weighting because such effects are treated in the precision of the individual T-values (through the weights w_i used in D^2).

For simplicity of notation we first consider only a simple sum of T-values, $\sum T_j$, where the index j identifies the different regions and/or seasons used. We need the standard error of this sum, and we then use the standard formula for the variance of a sum of correlated terms:

CPD
8, 263–320, 2012

$$\mathrm{Var}\left(\sum_j T_j\right) = \sum_j \mathrm{Var}(T_j) + 2\sum_{j_1<j_2} \mathrm{Cov}\left(T_{j_1}, T_{j_2}\right).$$

Thus, what we need together with the variances discussed in the previous section are the covariances. Consequently, we need to supplement the variance formula Eq. (10), by the corresponding formula for covariances, which can be written

$$\mathrm{Cov}\left(T_{j_1}, T_{j_2}\right) = \frac{1}{n^2}\left(\frac{1}{k} + \frac{1}{K}\right)\left\{2\,\mathrm{Cov}(\delta(j_1),\delta(j_2))^2 \sum_1^n w_i(j_1)\, w_i(j_2) + 4\,\mathrm{Cov}(\delta(j_1), \delta(j_2)) \sum_1^n w_i(j_1)\, w_i\,(j_2)(z_i(j_1) - \mu_x(j_1))(z_i(j_2) - \mu_x(j_2))\right\}. \quad (11)$$

Here $\mathrm{Cov}(\delta_i(j_1), \delta_i(j_2)) =,\rho(j_1, j_2)\sigma_{\delta(j_1)}\sigma_{\delta(j_2)}$, where ρ is the correlation coefficient. We have assumed that not only is the variance σ_δ^2 the same over time, as in formula Eq. (10), but that this also holds for the corresponding covariances. Note that the first term in the sum contains a covariance squared, corresponding to s_δ^4 in the variance formula Eq. (10). We assume that the covariances and the mean values $\mu_x(j)$ are estimated as with the variance σ_δ^2 and the μ_x in Eq. (10).

Now let nonequal weights be allowed, in the form $\sum_j c_j T_j$, where c_j are fixed coefficients which need not sum to 1. To express the corresponding calculations in this case, we arrange the variances and covariances for T_j in the covariance matrix $\mathbf{V}(\boldsymbol{T})$ for the vector $\boldsymbol{T}$ with components T_j. Let c be the corresponding column vector with components c_j. Then the variance for $\sum_j c_j T_j$ is obtained as the scalar

$$\mathrm{Var}\left(\sum_j c_j T_j\right) = c^T\, \mathbf{V}(\boldsymbol{T})\, c.$$

CPD
8, 263–320, 2012

We now have all requisites to calculate a unified performance metric, U_T, for each climate model under consideration:

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

Thus our final model score is a normalized sum of (possibly weighted) individual T-values for all available regions/seasons with proxy data, normalized by its standard error. This means that we can interpret U_T as a unified normalized test statistic of the null hypothesis, H_0, in the same way as for the individual T-values in the previous section. Hence, U_T can have a double usage; (i) to test if a forced climate model is better than unforced models, and (ii) as a rank value to compare different forced models; the more negative U_T-value the better (note that a forced model with $U_T > 0$ performs worse than the unforced models).

At this point, some practical issues are considered. In reality, the different proxy series may be of different lengths. This gives us reason to think of what n represents; recall that n is used in the calculation of individual D^2 values, and in the Var(T) and Cov(T) values. How should we choose n in the different parts of the calculations when the proxy records are of different length? We suggest to let the longest record determine n in all calculations. A consequence of letting the longest proxy series determine n is that more weight will be given to regional/seasonal data with long proxy series than those with short series, which seems reasonable. Note, however, that the mean values for the z, x_f and x_u series should be calculated only over the set of data available for the series in question (see Sect. 6).

Note that, in the period when all proxies are available, the weighting will be made both according to the proxy quality (through their respective w_i) and according to the variances and covariances of the T-values (which includes information from the behaviour of the simulated climate in the unforced models). If the additional weights c_j in the sum of T are used, then this will give further weighting to the data. We will,

CPD
8, 263–320, 2012

however, not discuss here how to construct such additional weights because we think this has to be determined uniquely for each particular set of available proxy data by external considerations, and no simple general rule seems plausible. In our pseudo-proxy experiment in Sect. 9, we will simply use equal weights (c_j).

8 Correlation as test statistic

As pointed out in Sect. 2, before any distance based performance metric is computed, one should first test if a forced climate model simulation is able to explain with statistical significance some part of the variation in instrumental and proxy data. If a forced climate model is unable to explain any variation in the instrumental and proxy data, then the D^2 and U_T measures provide little interpretable information. Here we suggest a test statistic, U_R, based on the correlation between a climate model simulation and the observations.

The x and z series are uncorrelated under H_0 (defined below), and (positively) correlated due to forcing effects appearing in both model simulations and real climate data. The stronger the forcing effect is in the model, the higher the expected correlation coefficient. We first consider a local test for a single grid-box (season) and next extend to a combination of data from several regions (and/or seasons).

We will again use notation z for the instrumental/proxy series, and the number of time units possible will be denoted n. For a particular grid-box (season), data may be available only during a shorter period of time, but with a weight factor that is zero when data is missing, as before, we can let n be the same for all grid-boxes (seasons).

The null hypothesis to be tested is:

H_0: *The climate model under consideration does not explain any of the temporal variation in the actual instrumental/proxy data.*

CPD
8, 263–320, 2012

Under H_0, we should of course not expect any significant correlation or covariance. However, unforced model simulations are important in providing a check that the test works reasonably under H_0.

We propose the following regression type statistic

$$R(x, z) = \frac{\sum w_i (\overline{x}_i - \mu_x)(z_i - \mu_z)}{\sum w_i^2 (z_i - \mu_z)^2}$$

for a given z series. We allow replicates of the same type of forced model, and we use their mean ($\overline{x}_i$) above. If only one replicate is available (or if only one replicate is tested), then $\overline{x}_i$ represents a single simulation. When $R(x, z)$ is normalized (divided) by its standard error, that is the square-root of its variance

$$\mathrm{Var}(R(x, z)) = \frac{(1/k)\,\sigma_\delta^2}{\sum w_i^2 (z_i - \mu_z)^2},$$

we get the correlation coefficient in a semi-empirical form, which is our test statistic for a single grid-box (season). As before, k is the number of replicates used to form $\overline{x}_i$, and the variance factor σ_δ^2 is again estimated from all available control runs, which we know satisfy the hypothesis H_0. The mean value μ_z is naturally estimated by the weighted average, $\overline{z} = \sum w_i z_i / \sum w_i$.

The weight factor w_i, however, is *not* the same weight factor as used with D^2 and T, because now only properties of the z series influence the weight. The principle is that the statistics $(z_i - \mu_z)$ should be weighted such that they get the same variance for all time units i. The weights should then be the following:

1. If $y = \tau$ (in periods where instrumental data with none, or negligible, noise is used): $w = 1$.

2. If $y = \tau + \theta$ (instrumental data with non-negligible noise variance, variance proportion q): $w = 1 - q$.

CPD
8, 263–320, 2012

3. If $y = \tau$, $z = \tau + \epsilon$ (proxy data are used, no noise in y, calibration period yields $\rho^2(y, z)$): $w = \rho^2(y, z)$.

4. If $y = \tau + \theta$, $z = \tau + \epsilon$ (proxy data are used, noise in y, calibration period yields $\rho^2(y, z)$): $w = \rho^2(y, z)/(1 - q)$.

Short term autocorrelation being present in unforced x series is avoided by the use of a sufficiently long time unit, as before. Short or long term autocorrelation in the x series due to modelled forcings will not be present under H_0 and therefore does not affect the validity of the test. On the contrary, we expect they are related to the actual variation in the z series and therefore will yield a significant test result.

With a number of grid-boxes (seasons) we assume, as for the test statistic T, that we form $\sum_j c_j R_j$ for some suitable coefficients c_j. To this end we need the variance for $\sum_j c_j R_j$. The variance for the local statistic R_j was given above, but we will also need the covariance between two such statistics. Given z, the covariance between R_{j_1} and R_{j_2} is given by the formula

$$\mathrm{Cov}\left(R_{j_1}, R_{j_2}\right) = \frac{(1/k)\rho(\delta_1, \delta_2)\sigma_{\delta_1}\sigma_{\delta_2}(1/n)\sum w_{1i}\, w_{2i}\left(z_{1i} - \mu_{z_1}\right)\left(z_{2i} - \mu_{z_2}\right)}{(1/n)\sum w_{1i}^2\left(z_{1i} - \mu_{z_1}\right)^2 \sum w_{2i}^2\left(z_{2i} - \mu_{z_2}\right)^2}.$$

Here ρ is the coefficient of correlation between the two x-sequences (from unforced simulations).

Finally, we arrange the variances and covariances for R_j in the covariance matrix $\mathbf{V}(\boldsymbol{R})$ for the corresponding vector $\boldsymbol{R}$. Let c be the corresponding column vector with components c_j. Then the variance for $\sum_j c_j R_j$ is obtained as the scalar

$$\mathrm{Var}\left(\sum_j c_j\, R_j\right) = c^T\, \mathbf{V}(\boldsymbol{R})\, c.$$

CPD
8, 263–320, 2012

Thus, our unified correlation-based test statistic thus becomes:

$$U_R = \frac{\sum_j c_j R_j}{\sqrt{\mathrm{Var}\left(\sum_j c_j R_j\right)}}.$$

Significant positive values of U_R mean that a model is able to explain some of the observed temperature variations. As with U_T, we can use the normal distribution to test for significance. Thus, for example, if $U_R > 1.65$ then H_0 can be rejected at the 5 % significance level. The larger the positive U_R values are, the stronger the correlation between the model and the observations. Any significant negative U_R would imply a negative correlation between a model and the observations. If such values are found, this may be a warning signal of possible erroneous calculations or possible problems with the climate model or proxy data. As with the T statistic, one may also consider a randomized significance test of R (see Appendix B).

9 A pseudo-proxy study addressing the detectability of solar forcing in simulated temperatures

9.1 Solar forcing

Variations of solar irradiance on long time-scales have not been directly measured but are potentially of great influence on global climate. However the amplitude of solar variability itself and the physical mechanisms involved in the global climate response to solar forcing are not yet fully understood (Gray et al., 2010). With the help of simple physical models, there have been recent attempts to reconstruct past solar irradiance over the last millennium using proxy information, such as sunspot observations (Krivova et al., 2003; Wenzler et al., 2006). However, there has been some discussion over the degree of background variation in irradiance possible on multi-decadal time-scales (Krivova et al., 2007). Many of the solar proxy time-series are not usable

CPD
8, 263–320, 2012

during the modern era of instrumental satellite observation due to recent contamination, such as by ^{14}C due to anthropogenic greenhouse gas emissions and nuclear testing (Schmidt et al., 2011). Shapiro et al. (2011) have questioned the instrumental reconstructions as well, with the satellite total solar irradiance (TSI) record being susceptible to systematic effects and instrumental wear.

The greatest current debate regarding uncertainty in solar variability over the past millennium centres on its magnitude (Ammann et al., 2007). Recent estimates of the change in solar radiative forcing (ΔF_{P-m}) from the Maunder Minimum period (late 17th century) to the present have mostly been in the range 0.1–0.2 $W\,m^{-2}$ (Krivova et al., 2007; Tapping et al., 2009; Steinhilber et al., 2009). Whilst potential magnifying processes have not been identified, the influence of solar forcing on climate remains uncertain (Wang and Sheeley, 2003; Hegerl et al., 2007b). For example, a recent analysis by Shapiro et al. (2011) suggests a $\Delta F_{P-m} = 1 \pm 0.5\,W\,m^{-2}$, entirely contrary to many recent estimates. When comparing a climate energy balance model (EBM) subjected to changing radiative forcing with multi-proxy temperature reconstructions, Friend (2011) concluded that it is quite possible that the amplitude of solar forcing is presently underestimated. Certainly the estimate of Shapiro et al. (2011) yields a far larger ΔF_{P-m} than any other recently published and adds to the debate regarding the magnitude of solar variability. Until new understanding of the physical mechanisms that govern solar variability progress, we have many different approaches to its reconstruction over the last millennium yielding different forcing values (Schmidt et al., 2011).

Given the uncertainty of solar forcing on the climate system on multi-decadal timescales, it is instructive to investigate its impact on climate models (Zorita et al., 2004; Ammann et al., 2007; Servonnat et al., 2010; Swingedouw et al., 2011). On sub-millennial time-scales the forced temperature variability during the pre-industrial period is thought to be dominated by atmospheric aerosol changes related to volcanic activity, as well as changes to solar energy output (Ammann et al., 2007). Zorita et al. (2004) and Swingedouw et al. (2011) have both linked a weakening of the Atlantic Meridional Overturning Circulation (AMOC) to higher solar forcing, an important internal

CPD
8, 263–320, 2012

dynamical system of the climate on multi-decadal time-scales. They also found an association with the North Atlantic Oscillation (NAO). Servonnat et al. (2010) investigated the importance of spatial scale on the degree of dominance regarding key climate forcings. They calculated a "signal to noise" ratio by dividing the globe into N regions, namely the ratio of the variances of N temperature time-series in a forced simulation (solar, greenhouse gas, insolation) against the corresponding time-series variances of a control simulation. They found the characteristic spatial scale suitable for detecting a forced signal in their model to be approximately 5×10^6 km, an area roughly the size of Europe. However, this was for a single climate simulation and not an ensemble.

The uncertainty in the magnitude of solar radiative forcing and its impact on the climate system is a greater source of uncertainty than that associated with the timing of solar activity variations over the last millennium (Muscheler et al., 2007). It is the intention of the present analysis to use the developed statistical framework to rank or distinguish between model simulations using a variety of forcings, either as individual forcings added to a control model or several used in tandem. Given the uncertainty already described in solar forcing over the last millennium, it is important to compare simulations using alternate forcing histories rather than selecting arbitrary series (Schmidt et al., 2011). Hence, a suitable set of simulations were used from the Community Earth System Modeling (COSMOS) Millennium Activity of the Max Planck Institute (Jungclaus et al., 2010). Here two solar radiative forcing time-series were employed of differing magnitude, as well as principal drivers of pre-industrial climate over the last millennium (orbital, solar, volcanic as well as land-cover changes – Hegerl et al., 2007a; Crowley, 2000). The solar, volcanic and land-use change forcings were used individually and jointly in the COSMOS Millennium Activity and these simulations were analyzed here. It is hoped that the developed methods can be used to distinguish between these forced climate simulations in order to deduce the most important climate forcings. This will allow better judgement regarding how possible it is, in future comparisons, to identify which simulation is best able to simulate observed temperatures in real proxy and instrumental data.

CPD
8, 263–320, 2012

9.2 The COSMOS Millennium Activity – model description and experimental design

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

The project framework involved the creation of a 3000-year unforced control (CTRL) simulation, after a multi-century spin-up phase in which the carbon cycle was brought into equilibrium. This CTRL model experienced 800 AD orbital conditions and pre-industrial greenhouse gas concentrations (Jungclaus et al., 2010). The various forcing reconstructions were added individually or in combination to the CTRL reference boundary conditions. To account for the uncertainty in solar forcing, the TSI forcing used in this project involves both a "low" (or standard) forcing exhibiting a total increase of 0.1% from the Maunder Minimum to the present day (in agreement with contemporary evaluations during the model setup; Krivova et al., 2007; Steinhilber et al., 2009; Tapping et al., 2009) against a forcing with a "high" amplitude reduction in the Maunder Minimum (0.25 %) compared with the present (Bard et al., 2000).

The globally averaged land-only annual temperature anomalies (30-year means) of the COSMOS simulations are shown in Fig. 1. The simulations have been distinguished into the CTRL (itself separated into three millennial sections) in Fig. 1a, the single forcing simulations (Fig. 1c) and two full-forcing ensembles E1 (Fig. 1b) and E2 (Fig. 1d). The evolution of the forcings applied to the models are shown in Fig. 2. A representation of the forcings is shown in Fig. 2, consisting of atmospheric CO_2 concentrations, land-use changes, volcanic as well as two separate solar forcing curves,

CPD
8, 263–320, 2012

the “high” and “low” solar series from Bard et al. (2000) and Krivova et al. (2007) respectively. The two full-forcing ensembles were generated by initiating different ocean boundary conditions and are separated by their respective “high” (E2) and “low” (E1) solar forcing and any solar-induced CO_2 concentration changes (made possible by the interactive carbon cycle model).

Figure 1c shows the four single forcing simulations: land-use changes (green), low solar (light orange), high solar (yellow) and volcanoes (dark orange). The high solar and volcanic simulations exhibit some multi-decadal variability, such as the low temperatures in the volcanic series during the late 13th century or the high temperatures in the high solar simulation in the late 18th century. In contrast, the land-use and low solar simulations exhibit less variability on the multi-decadal to centennial time-scales. In Fig. 1b, the E1 ensemble members feature generally warmer conditions from 800 AD until a large cooling occurs in the latter half of the 13th century. Conditions are thereafter generally cooler, including a cold period during the 19th century, until a warming occurs in the 20th century partly associated with CO_2 radiative forcing (Fig. 2a). The single forcing simulations do not show this 20th century warming as they do not contain the CO_2 radiative forcing, this was only included in the E1 and E2 ensemble simulations. In Fig. 1d, the E2 simulations exhibit larger variance than found for the E1 simulations, as can be expected when comparing the low and high solar radiative forcing histories (Fig. 2a). In particular, the E2 simulations display a more pronounced difference between the warm period from 800 AD to the late 13th century, compared with the generally cooler conditions during the latter half of the last millennium. The single time-series representations of the global forcings are shown in Fig. 2, in terms of their annual mean radiative forcing at the top of the atmosphere. The CO_2 radiative forcing is shown only for one of the E1 simulations as they are all similar in evolution, although not identical due to the carbon cycle feedback.

CPD
8, 263–320, 2012

9.3 Model – (pseudo-proxy) data comparison setup

In order to investigate the developed statistical methods, a pseudo-proxy analysis was conducted using the COSMOS simulations. A pseudo-proxy series can be defined as any instrumental or climate model data that has been distorted through the addition of noise (Jones et al., 2009). This is to ensure that the pseudo-proxies account for a fraction of the variance of a temperature series, as is the case for a real proxy reconstruction of temperature. A key advantage of this approach is that the distortion and reconstruction target are both prescribed and hence fully known. In the present pseudo-proxy analysis, τ_i is defined explicitly by a particular simulation and is chosen as the "true climate" where the noise level and regions used in the comparison are specified. Then the proxy series z_i and instrumental series y_i can be defined as the "true climate" simulation plus added noise.

An advantage of the pseudo-proxy approach using model output is that the number of locations i can be varied from a single grid-box to any number of locations. We also consider an average single time-series for the entire globe. We use the annual mean temperature and land points only, as most real high-resolution proxy series are found on land. This analysis technique allows a unique insight into the proposed correlation and distance comparison methods, since when the true climate is defined as one of the full-forcing simulations, the solar (and other) radiative forcing amplitudes will be known. Given a realistic amount of noise in the pseudo-proxies, it is hoped that the two comparison methods will distinguish between the E1 and E2 ensemble simulations when a single member of one of those ensembles is used as the "truth". In other words, if a particular E1 member is used as the target and no significant distinction can be made between the E1 and E2 ensemble members based on their proximity to the chosen "true climate" simulation from E1 (with realistic noise added), then the method cannot be expected to help better constrain definition of a suitable past millennial solar forcing amplitude, if this analysis were applied to real proxy and instrumental data.

CPD
8, 263–320, 2012

The climate model simulation time sequences x_i are taken from the COSMOS simulations as illustrated in Fig. 1, where the forced component $\alpha\xi_i$ is the response to either a single forcing in the case of land-use changes, solar and volcanic, or to the combined forcings in the E1/E2 ensembles. $\alpha = 0$ in the CTRL models and x_i is from statistical model 2. As previously stated, the true temperature τ_i is known and given as the target series in question, by selecting one of the E1 or E2 ensemble member simulations. The instrumental measurements y_i are defined as the target simulation for a given location over the period 1850–2000 with added white noise (θ_i), defined as having 10 % of the variance of y_i. Regarding the added noise in y_i, this approximately corresponds to a doubling of recent single-thermometer measurement error estimates (Folland et al., 2001; Brohan et al., 2006). The proxy series z_i are defined similarly, though over the period 1000–2000 and featuring added white noise with 2/3rds the variance of z_i. This corresponds to an SNR = 0.71 (signal-to-noise ratio) and correlation $r = 0.58$ between z_i and τ_i, as is the case for many high-quality real proxy records. Note that both higher and lower percentages were also investigated (see electronic supplement).

The analysis was conducted on the 1000-year period, 1000–2000 AD (although the forced simulations begin at 850 AD) in order to separate the 3000-year unforced simulation into three 1000-year control simulations as to be used in the comparison. The computation of U_T and U_R, however, was restricted to the period 1000–1850 to avoid the influence of anthropogenic greenhouse gas increases. Data after 1850 were used only in the calibration of z_i against y_i and for estimating the variance of y_i. The simulation time evolutions shown in Fig. 1 are presented as they are used in this comparison, namely for land-only and for 30 year non-overlapping means. Land-only grid-boxes were used in this analysis, to reflect the reality of the fact that annually resolved proxy data from the oceans are rare. A motivation for using 30-year means is given below.

Recall from Sect. 2 that the unforced simulated temperature δ_i is assumed to be white noise. It is of course quite possible that white noise is not a good representation of the internal variability in the true climate, but as a distance measure D^2 does not require white noise. However, the null hypothesis of the statistical tests is that forced

CPD
8, 263–320, 2012

simulations are equivalent with unforced (CTRL) simulations, so for the described tests to have the prescribed type I error level, the unforced simulations should be well represented by white noise. Short term memory of model climate systems has been questioned (Cohn and Lins, 2005; Rybski et al., 2008). We investigated the seriousness of this problem by calculating the lag-1 autocorrelation for the 3000-year CTRL simulation, both in terms of the proportion of global area with significant autocorrelations for various time-resolutions, as well as the lag-1 autocorrelation for the global land-only series (see electronic supplement for further details). It was found that beyond a time-resolution of 20 years, δ_i can be considered as white noise, in keeping with the statistical assumptions of Sect. 2. Hence a non-overlapping 30-year mean resolution, as used in the present analysis, should be able to keep the type I error of the tests under reasonable control in the model – (pseudo-proxy) data comparisons undertaken here.

9.4 Model – (pseudo-proxy) data comparison

We first conducted a study for global average (area-weighted) time-series using only land points (Fig. 1), the results of which are shown in Fig. 3. The global mean was investigated first, simply because this series will likely exhibit a high SNR of the forced component to natural internal variability (Servonnat et al., 2010). Hence we use a single set of τ, y and z series in this globally averaged analysis. For the analysis in Fig. 3, a SNR of 0.71 is prescribed (i.e. the noise variance in z is two-thirds of the total variance of z). Both the E1 and E2 simulations were used separately as targets in this experiment, and to use as many target "true" climates or "truths" as possible, each member was used as the target in turn.

For each type of "truth", ≈100 noise realizations were generated to produce y and z with a rotation in the five E1 target simulations (20 noise realizations for each simulation, $5 \times 20 = 100$) (Fig. 3a and c) and in the three E2 target simulations (33 noise realizations for each simulation, $3 \times 33 = 99$) (Fig. 3b and d). Recall that the difference

CPD
8, 263–320, 2012

between the E1 and E2 ensembles is the use of the "low" and "high" solar forcing series respectively. Comparing the different ensemble members as targets should enable us to draw conclusions about how useful the correlation and distance measures are for ranking the simulations. Here, these measures can be judged in terms of their ability to distinguish the temperature response to different solar forcings, given the presence of volcanic and land-use forcing in the target simulations as well as internal variability. Note that there is in fact also a subtle change to the atmospheric greenhouse gas concentrations in the model for each full-forcing ensemble member but this difference is thought to be of no importance, as the forcing time-series are almost identical and of little importance before 1850. Iteratively treating the E1 or E2 ensemble members as targets could cause the distributions to be hierarchical, where the distribution of U_T and U_R for different noise iterations could potentially be non-overlapping when comparing different ensemble members. An identical analysis to this was conducted, with zero noise added to the target, which revealed the E1 and E2 ensemble simulations to give very similar results with little qualitative spread (*not shown*). This satisfied the authors sufficiently that the spread of the distributions in Fig. 3 predominantly represent the uncertainty due to the pseudo-proxy noise realizations, whereas the locations of the medians of the distributions largely represent the effect when comparing simulations from different climate models.

To further explain the U_T and U_R box-plot distributions shown in Fig. 3, the first four represent the single forcing simulations, namely land-use changes (green), low solar (light orange), high solar (yellow) and volcanoes (dark orange), where they are compared with either the E1 (left panels) or the E2 (right panels) simulations as target. Analogously, the next five box-plots (numbers 5-9) represent the E1 simulations, all coloured light blue with their corresponding ensemble average U_R/U_T value in dark blue (number 13). The three E2 simulations are coloured light red (numbers 10–12) with their corresponding ensemble average in red (number 14). Hence, both the individual E1/E2 simulations and the E1/E2 ensemble averages were compared with target simulations. Note that when an E1 (or E2) simulation is used as the target, this

CPD
8, 263–320, 2012

target simulation is excluded from the E1 (or E2) ensemble being analysed (i.e. being compared with the target).

Figure 3a and b show the U_R correlation analysis results, whilst Fig. 3c and d show the U_T distance measure results. Figure 3a and b also feature the three CTRL simulation segments (numbers 15–17) as these are not required in the calculation of U_R. From Fig. 3a, the U_R correlation analysis with E1 as target, it is clear that the individual E1 ensemble simulations (numbers 10–12 and 14) are significantly correlated with each other. However, the E2 simulations are the most highly correlated simulations with the E1 targets. This can be expected in so far as the E2 simulations feature the strongest solar forcing and the largest variability (Fig. 1). The volcanic forcing seems to explain much of the correlations for E1, looking at the significance of U_R for the volcanic simulation (number 4) when E1 serves as target (Fig. 3a). When E2 serves as target, both the volcanic *and* high solar (number 3) explain much of the E2 correlations (Fig. 3b). The land-use forcing simulation (number 1), as well as the low solar simulation (number 2), are not significantly correlated with either the E1 or E2 target ensembles. When using the E2 ensemble as target (Fig. 3b), both the E1 and E2 simulation members and ensemble averages are significantly correlated, which can be expected given, in particular, the shared volcanic forcing series.

Whilst it is clear from Fig. 3a and b that both the E1 and E2 ensembles are significantly correlated with each other (according to U_R), this will not necessarily be reflected in the distance measure U_T for these full-forcing simulations. The distance measure U_T (Fig. 3c and d) is expected to be more effective in distinguishing between the simulations and, in some instances, being capable of ranking them. When E1 serves as target (Fig. 3c) the E1 simulations (numbers 5–9) and their average (number 13) are mostly significantly closer to the target than the CTRL simulations, whereas the E2 simulations (numbers 10–12) and E2 ensemble average (number 14) are not. This differs from the U_R analysis using E1 as target (Fig. 3a) in that the E1 and E2 simulations can be distinguished using U_T. Note that both measures show similar results when E2 serves as target (comparing Fig. 3b and d). These results reflect that the U_T distance

CPD
8, 263–320, 2012

measure can sometimes distinguish between simulations with similar U_R values for a given target. The principal difference between these two methods is that the correlation analysis does not consider the variance of the two compared series (target and simulation), whereas this is explicitly considered in the distance measure. For both the U_R and U_T values the low solar simulation (number 2) is not significantly correlated with, or close to, the E1 targets (Fig. 3a and c). In contrast, the high solar simulation (number 3) is significantly correlated with, and close to, the E2 targets (Fig. 3b and d). This implies that the low solar forcing is too weak to produce any detectable effect at the 30-year time-scale, whilst the high solar *is* strong enough. A related conclusion was reached by Ammann et al. (2007): the greater the solar forcing amplitude applied to their model, the weaker the detectable response to other natural forcings.

On small spatial scales, the ability to distinguish between simulations that use low and high solar forcing and consequently rank them, may not be possible. At global or hemispheric scales, the temperature can be expected to respond to large-scale external forcings (such as solar or greenhouse gases), whereas at local or regional scales the internal climate dynamics can account for a larger proportion of the temperature variability (Goosse et al., 2005). Although looking at seasonal European climate change, Hegerl et al. (2011) were able to attribute external forcing as a contributing factor to changes in winter and summer temperature.

It is the case that despite ongoing attempts at reconstructing hemispheric or global temperature series (e.g. Moberg et al., 2005; Juckes et al., 2007; Mann et al., 2008), the coverage of millennial high quality proxy data is comparatively sparse, particularly in the southern hemisphere. Given the uncertainty in hemispheric or global reconstructions due to differing calibrations, reconstruction algorithms and proxy networks (Esper et al., 2005; Christiansen et al., 2009; Frank et al., 2010), it is useful to look at a realistic representation of the kinds of proxy data networks likely to be used with real proxy data.

In order to investigate how successfully the correlation or distance measures can be used to distinguish high and low solar simulations on a current example of a published

CPD
8, 263–320, 2012

global proxy network, the locations of 27 proxy series in Juckes et al. (2007) were used for the correlation and distance analysis (Fig. 4). White noise was added to the target series at each location such that the SNR = 0.71. Though this set of locations is clearly a sparse representation of the global surface, such a set is typical of where proxy data have been acquired in the past and 20–40 or so locations is a realistic number as could be used currently in a true model-data comparison for the last millennium.

Results for the same type of experiments conducted in Fig. 3 are shown in Fig. 5, but for the combined Juckes et al. (2007) locations. Specifically, we compute local correlation (R) and distance (D^2) measures for the 27 locations, before they are combined to obtain a single U_R and U_T value for each simulation. The correlation analysis U_R for the Juckes et al. (2007) proxy locations gives similar results to the global time-series analysis, though surprisingly the correlations are not less significant. This is something that could haven been expected due to the increased influence of internal (unforced) variability at the regional scale in combination with the reduced area coverage. As in the global average analysis, the E1 and E2 simulations are generally significantly correlated with both the E1 (Fig. 5a) and E2 (Fig. 5b) targets (with the exception of simulation number 8 in Fig. 5b). However, when E1 serves as target, the distance measure U_T is unable to distinguish the E1 simulations from the CTRL simulations (Fig. 5c), whereas the E2 simulations are significantly closer to the target than the CTRL simulations when E2 serves as target (Fig. 5d). Concerning the single forcing simulations, only the high solar (number 3) is significantly correlated with the E2 targets and closer to the target than the CTRL simulations (Fig. 5c and d).

Using a realistic set of proxy locations such as the Juckes et al. (2007) set, it seems there is no clear distinction between simulations, in terms of U_T ranking or scoring, unless the forcing is large and multi-decadal in nature (as is the case for the high solar forcing used here). Note that U_R is more sensitive than U_T for testing if a model forcing has any correspondence with the true climate, but it answers a different question than U_T. This higher sensitivity is seen when we compare subfigures a and b with c and d respectively in both Figs. 3 or 5. Specifically, if U_R is not significant, neither is U_T.

CPD
8, 263–320, 2012

Comparisons between the Juckes et al. (2007) and global land-only average results naturally lead to the question of how the possibility to rank simulations depends on the spatial coverage of the pseudo-proxy data.

9.5 Varying coverage

There are in practice relatively few locations which have high quality proxy data available or where there is the potential at present to acquire more data. A pseudo-proxy experiment, however, has the advantage of allowing any number of locations to be used to serve as a proxy series or instrumental series. Hence, an analysis is conducted on how varying degrees of % coverage affects the ability of the correlation and distance measures to distinguish between the high and low solar single forcing simulations and between the E1 and E2 ensembles, when either E1 or E2 serves as target.

The various global coverages are for 0.1, 0.25, 0.5, 1, 2, 3, 4, 5 %, using only land grid points, which is equivalent to 3, 10, 22, 44, 90, 137, 183, 230 proxy locations. 20–40 or so proxy locations is a typical number of high quality millennial proxy data found in current analyses (Christiansen and Ljungqvist, 2011). Calculation of the covariance matrices $\mathbf{Cov}(T_{j_1}, T_{j_2})$ and $\mathbf{Cov}(R_{j_1}, R_{j_2})$ becomes computationally intensive for large % coverages, hence they were only calculated up to 5 %. The set of proxy locations were selected as a stratified random sample from the available land points in the COSMOS simulations, with specified proportions for three strata (the latitudinal bands 0–30°, 30–60°, 60–90°). The stratification was chosen to better control the coverage and to account for the changing area of the grid points with latitude in the simulations.

Figure 6 shows the correlation U_R (top panel) and distance U_T (bottom panel) measures for the low (light orange) and high solar (yellow) single forcing simulations for different % coverages. As with the previous analyses, ≈100 noise realizations were generated for each coverage level. The filled lines represent the median values, whilst the upper and lower quartiles are dashed. For comparison, results for the volcanic (dark orange) forcing simulation are also shown. As with Figs. 3 and 5, the left panels (Fig. 6a and c) are with E1 as the target and the right panels (Fig. 6b and d) are with E2

CPD
8, 263–320, 2012

as target. The high solar simulation is significantly correlated even for the lowest coverages when E2 serves as target (Fig. 6b), whilst also achieving significant U_R values for coverages upwards of 1 % when E1 serves as target. The latter result was discussed in Sect. 9.4. Contrastingly, the high solar simulation U_T values are significantly better than the CTRL simulations when E2 serves as target (Fig. 6d), but not when E1 serves as target (Fig. 6c). The volcanic simulation is mostly significantly correlated with both E1 and E2 targets (Fig. 6a and b) and its U_T values are generally only significant for coverages upwards of 1 % for both targets (Fig. 6c and d). The low solar simulation shows no significant correlations for either target ensemble and can therefore be expected to be indistinguishable from the CTRL simulations using the U_T measure.

Figure 7 is as Fig. 6 but shows the E1 (blue) and E2 (red) ensemble average results for U_R and U_T. Both ensembles correlate with their own ensemble targets even at the lowest data coverages. As with the globally-averaged analysis of U_R (Fig. 3a and b), the E1 and E2 ensembles cannot be easily distinguished when using E1 as a target (Fig. 7a), but are distinguishable when E2 serves as target (Fig. 7b). The results for U_T are much the same as for the global analysis, where the E1 and E2 ensembles can be correctly ranked with their respective targets. Notably for coverages lower than 1 % it become difficult to distinguish E1 from the CTRL simulations or separate the E1 and E2 simulations when E1 serves as target (Fig. 7c). When E2 serves as target, it is always easy to distinguish between and correctly rank (when using U_T) the E1 and E2 ensembles. Additionally, the experiments of Figs. 6 and 7 were conducted for cases with a SNR = 0.25 and also with negligible noise, the results of which are briefly discussed in the conclusions and shown in the electronic supplement. An important feature of Figs. 6 and 7 to note is how flat the U_R and U_T measures are with changing % coverage. In fact, there is little gain in increasing the sample size from 20–40 or so proxy series to several hundred. Above all else, this suggests a substantial degree of spatial correlation in temperature, given the 30-year time-resolution used in this analysis (Jones et al., 1997; Franke et al., 2011).

CPD
8, 263–320, 2012

10 Conclusions

In view of an increasing need to be able to compare output from model simulations with climate variables reconstructed from proxy records, a method has been developed for comparing temperature on a variety of time and spatial scales. The methodology developed here was specifically designed in view of the irregular distribution of such climatic records, with differing representativeness concerning seasons and record length. We provide a statistical framework for these comparisons explicitly using information on proxy uncertainties based on their correlation with instrumental data, which allows the assessment of which simulations are statistically similar to a target series.

Specifically, two goodness-of-fit measures have been developed; a unified correlation-based test statistic U_R as well as a distance-based test statistic U_T. Given the ongoing uncertainty in solar variability over the last millennium it is hoped that the latter of these measures could be used to detect different temperature responses in simulations with two different solar forcings over this period, despite there being noise in the target data, as well as internal and forced variability in the climate system. A pseudo-proxy experiment was designed for this task, based on the MPI-COSMOS earth system model simulations (Jungclaus et al., 2010). The advantage of a pseudo-proxy experiment is that the true climate or "truth" is always known and the statistics of the added noise in the proxies are also known. Hence if no difference between two forced simulations containing different solar forcing evolutions can be detected with these methods, then no significant conclusions could be assumed based on comparing model output with real proxy data. The E1 and E2 ensembles provided a useful 800–2005 AD time interval, both containing forcing information from land-use changes, volcanic aerosols as well as greenhouse gas concentrations; but with different solar forcing series. Namely, E1 features a low solar forcing variability (0.1 % increase from the Maunder Minimum to the present climate) whilst E2 features a stronger solar variability (0.25 %).

CPD
8, 263–320, 2012

Firstly an analysis was conducted on globally averaged land-only data as this was thought to contain the highest SNR in regards to forced- against internal variability. Hence a single series was calculated for each simulation and compared with every member of the E1 and E2 ensembles in turn plus added noise. When E1 serves as a target, it was found that both E1 *and* E2 simulations were strongly correlated with the targets, both having similar and significant U_R values. This suggests that the shared forcing information gives significantly correlated temperature evolutions in both E1 and E2 simulations. However, when the U_T statistics are viewed with E1 as target, the two ensembles can be distinguished. In other words, the different simulation ensembles can be ranked based on their different solar forcing using U_T. This is largely due to the fact that U_T accounts for the variance of the two compared series, whereas U_R does not. When E2 serves as target, the stronger solar forcing used in this ensemble renders it detectably different from the E1 ensemble in terms of both U_R and U_T statistics. However only the U_T measure should be used to rank the simulations.

Given that this global comparison is hypothetical due to the heterogeneous nature of real proxy data coverage in space and time, and that this statistical framework has been developed in view of using real proxy information to assess the goodness-of-fit of model simulations, a representative set of proxy locations was taken from Juckes et al. (2007) to conduct the pseudo-proxy comparison on. The results of which were similar to the global land-only analysis, however, the U_T values of the E1 ensemble could not be said to be significantly different from the CTRL simulations when E1 serves as target. This motivated an analysis of how differing % coverage levels change the detectability of significance in the U_T and U_R statistics (Figs. 6 and 7). Additionally, the same type of analysis was conducted for a higher noise level (SNR = 0.25) as well as with negligible noise (see electronic supplement). For an SNR = 0.25 the U_R and U_T of the E1 and E2 ensemble simulations are indistinguishable when E1 serves as target, whilst with E2 as target and for coverages above 0.5 %, they are still distinguishable. Reducing the noise to negligible levels allows distinction between the E1 and E2 ensemble simulations with E1 serving as target beyond 0.5 % coverage, however this is an extremely hypothetical

CPD
8, 263–320, 2012

case where the "true" climate is represented near perfectly by the available target data. Nevertheless, these results suggest for a global coverage of say 40 or more proxy locations, if a very high quality of individual proxy series is obtained with low noise levels, it can be possible to distinguish the E1 and E2 ensembles when using E1 as target. If E2 serves as target very few proxy series are needed, even less than 10 given a SNR of at least 0.71. These results have an important implication: it is more important to improve the quality of individual proxy series in terms of SNR than it is to increase the quantity of available proxies.

There are additional methodological considerations for the future development of these test statistics, such as the allowance of more realistic noise types (e.g. red noise) or comparisons over different time resolutions and periods. Additionally, the aim is that future analyses can be conducted using these test statistics to compare model simulation output with real proxy data. By varying the SNR and coverage levels in the pseudo-proxy section of this analysis (Sect. 9), we conclude that in order to successfully rank model simulations and be able to draw conclusions about past climate forcing series, it will be necessary to gain more very high quality proxy series as to add to an analysis. That is not to say there is no value in obtaining proxy data with low SNR values regarding a particular climate variable. Individual proxy records can provide useful corroborative information when developing regional syntheses or interesting holistically consistent documents of local environmental change. However from the results of the present analysis, it seems the benefits of adding many proxy series with low SNR values, are negligible in comparison to improving SNR values in 20–40 or so proxy series.

CPD
8, 263–320, 2012

Appendix A

Averaging inside D^2

Here we provide the necessary formulae for calculating the bias correction and for estimating the variance of T when using the difference $T(\overline{x_f}, \overline{x_u}, z) = D_w^2(\overline{x_f}, z - D_w^2(\overline{x_u}, z)$ as the distance-based test statistic, i.e. with averaging inside D^2.

A1 Bias of the test statistic T

Suppose x_f includes a forced component $\alpha\xi$. When

$$T(x_f, x_u, z) = \overline{D_w^2}(x_f, z) - \overline{D_w^2}(x_u, z),$$

that is under outside averaging, the expected value of T is

$$E(T) = -\left(2\alpha - \alpha^2\right)\frac{1}{n}\sum_{i=1}^{n} w_i(\xi_i - \mu)^2.$$

Under H_0, $\alpha = 0$ and the expected value $E(T)$ is zero.

With

$$T\left(\overline{x_f}, \overline{x_u}, z\right) = D_w^2\left(\overline{x_f}, z\right) - D_w^2\left(\overline{x_u}, z\right),$$

that is under inside averaging, the expected value of T contains an additional bias term, and is

$$E(T) = -\left(2\alpha - \alpha^2\right)\frac{1}{n}\sum_{i=1}^{n} w_i(\xi_i - \mu)^2 + \sigma_\delta^2\left(\frac{1}{k} - \frac{1}{K}\right)\frac{1}{n}\sum_{i=1}^{n} w_i.$$

The bias term is zero only when $k = K$. Thus, if inside averaging is used with $k \neq K$, the bias must either be judged negligible, or estimated and corrected for.

CPD
8, 263–320, 2012

A2 Precision of the test statistic T

Under inside averaging, the analytical formula for the variance of T, under assumed normality of $x_f - x_u$ and given the z sequence, is:

$$\mathrm{Var}\left(T\left(\overline{x_f}, \overline{x_u}, z\right)\right) = \frac{2\,\sigma_\delta^4}{n^2}\left(\frac{1}{k^2} + \frac{1}{K^2}\right)\sum_1^n w_i^2 + \frac{4\,\sigma_\delta^2}{n^2}\left(\frac{1}{k} + \frac{1}{K}\right)\sum_1^n w_i^2 (z_i - \mu_x)^2.$$

A3 Results

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

Appendix B

Alternative reference distributions for D_w^2, T and R

It deserves mention that there are (at least) two possible nonparametric alternatives to the normality-based tests for H_0 used above, instead being based on exchangeability, either between replicated simulations x_u or different time intervals within simulations x_u of the unforced model. First, if the number K of available unforced simulations is large, and $k \ll K$, then we could repeatedly take random samples of size k out of the K, to let

CPD
8, 263–320, 2012

them represent x_f under H_0. Along these lines a reference distribution for the distance measure D_w^2 or the D^2-difference T could be estimated, valid under H_0. However, to have a large number of unforced simulations (say 50) using a single climate model does not appear to be realistic at present. Presumably a better alternative is to utilize the stronger property of exchangeability within an unforced sequence. This assumes negligible autocorrelation, to be accomplished by a large enough time unit, see Sect. 6. From each sequence x_u, new sequences can be generated by randomly permuting the order within the sequence. For each such new sequence the corresponding D_w^2 and T values can be computed, and this leads to a reference distribution for D_w^2 or T under H_0. In the present study such random permutation-based tests have not been applied, but the aim is to try them in later studies (in progress). However, it should not be forgotten that the primary use of D_w^2 is for ranking different simulations, and for that purpose the reference distribution of the test statistics is of somewhat limited interest, and the more explicit formula U_T proposed in Sect. 7 appears to be more convenient.

Analogous constructions can be used for the correlation measure R. A reference distribution could be constructed by computing R for each of a large number of replicated simulations. If only one or a few simulations are available, we are confined to running through random permutations of their time order before correlating them with the instrumental/proxy sequence.

Supplementary material related to this article is available online at: http://www.clim-past-discuss.net/8/263/2012/cpd-8-263-2012-supplement.pdf.

Acknowledgements. A. M. coordinated the study. R. S. developed the statistical framework. A. H. wrote the computer code and performed the numerical analyses. The paper was written jointly. We thank the Swedish Research Council (grants 70454201 and 90751501) and the European Union (FP6 grant 017008, "Millennium" project) for funding. We also thank Johann Jungclaus of the Max Planck Institute for providing the COSMOS data as well as help and advice regarding the simulations.

References

Allen, M. R. and Tett, S. F. B.: Checking for model consistency in optimal fingerprinting, Clim. Dynam., 15, 419–434, 1999. 265

Ammann, C. M., Joos, F., Schimel, D. S., Otto-Bliesner, B. L., and Tomas, R. A.: Solar influence on climate during the past millennium: Results from transient simulations with the NCAR Climate System Model, P. Natl. Acad. Sci., 104, 3713–3718, 2007. 289, 298

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. 291, 292

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, Journal of Geophysical Research, 111, D12106, doi:10.1029/2005JD006548, 1–12, 2006. 294

Brown, P. J.: Measurement, Regression and Calibration, Oxford University Press, Oxford, UK, 1993. 275

Christiansen, B. and Ljungqvist, F. C.: Reconstruction of the extra-tropical NH mean temperature over the last millennium with a method that preserves low-frequency variability, J. Climate, 24, 6013–6034, 2011. 300

Christiansen, B., Schmith, T., and Thejll, P.: A Surrogate Ensemble Study of Climate Reconstruction Methods: Stochasticity and Robustness, J. Climate, 22, 951–976, 2009. 298

Cohn, T. A. and Lins, H. F.: Nature's style: Naturally trendy, Geophys. Res. Lett., 32, L23402, doi:10.1029/2005GL024476, 2005. 295

Crowley, T.: Causes of climate change over the past 1000 years, Science, 289, 270–277, 2000. 290

Esper, J., Wilson, R. J. S., Frank, D. C., Moberg, A., Wanner, H., and Luterbacher, J.: Climate: past ranges and future changes, Quaternary Sci. Rev., 24, 2164–2166, 2005. 298

Folland, C. K., Rayner, N. A., Brown, S. J., Smith, T. M., Shen, S. S. P., Parker, D. E., Macadam, I., Jones, P. D., Jones, R. N., Nicholls, N., and Sexton, D. M. H.: Global temperature change and its uncertainties since 1861, Geophys. Res. Lett., 28, 2621–2624, 2001. 294

Frank, D. C., Raible, C. C., Büntgen, U., Trouet, V., and Stocker, B.: Ensemble reconstruction constraints on the global carbon cycle sensitivity to climate, Nature, 463, 527–530, 2010. 298

CPD
8, 263–320, 2012

Franke, J., Gonzalez-Rouco, J. F., Frank, D., and Graham, N. E.: 200 years of European temperature variability: insights from and tests of the proxy surrogate reconstruction analog method, Clim. Dynam., 37, 133–150, 2011. 301

Friend, A. D.: Response of Earth's surface temperature to radiative forcing over A.D. 1-2009, J. Geophys. Res., 116, D13112, doi:10.1029/2010JD015143, 2011. 289

Goosse, H., Renssen, H., and Bradley, R. S.: Internal and forced climate variability during the last millennium: a model-data comparison using ensemble simulations, Quaternary Sci. Rev., 24, 1345–1360, 2005. 272, 298

Goosse, H., Renssen, H., Timmermann, A., Bradley, R. S., and Mann, M. E.: Using paleoclimate proxy-data to select optimal realisations in an ensemble of simulations of the climate of the past millennium, Clim. Dynam., 27, 165–184, 2006. 266, 272

Graham, N. E., Ammann, C. M., Fleitmann, D., Cobb, K. M., and Luterbacher, J.: Support for global climate reorganization during the "Medieval Climate Anomaly", Clim. Dynam., 37, 1217–1245, 2011. 265

Gray, L. J., Beer, J., Geller, M., Haigh, J. D., Lockwood, M., Matthes, K., Cubasch, U., Fleitmann, D., Harrison, G., Hood, L., Luterbacher, J., Meehl, G. A., Shindell, D., van Geel, B., and White, W.: Solar influences on climate, Reviews of Geophysics, 48, RG4001, doi:10.1029/2009RG000282, 2010. 288

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, 2007a. 265, 290

Hegerl, G. C., Zwiers, F. W., Braconnot, P., Gillett, N. P., Luo, Y., Marengo Orsini, J. A., Nicholls, N., and Penner, J. E., and Stott, P. A.: Understanding and Attributing Climate Change. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., and Miller, H. L., Cambridge University Press, Cambridge, UK and New York, NY, USA, 2007b. 265, 289

Hegerl, G. C., Luterbacher, J., Gonzalez-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. 265, 298

Jones, P. D., Osborn, T. J., and Briffa, K. R.: Estimating Sampling Errors in Large-Scale Temperature Averages, P. Natl. Acad. Sci., 10, 2548–2568, 1997. 301

CPD
8, 263–320, 2012

Jones, P. D., Briffa, K. R., Osborn, T. J., Lough, J. M., van Ommen, T. D., Vinther, B M., Luterbacher, J. W. E. R. Z. F. W., Mann, M. E., Schmidt, G. A., Ammann, C. M., Buckley, B M., Cobb, K. M., Esper, J., Goosse, H., Graham, N., Janse, E., Kiefer, T., Kull, C., Küttel, M., Mosley-Thompson, E., Overpeck, J. T., Riedwyl, N., Schulz, M., Tudhope, A. W., Villalba, R., Wanner, H., Wolff, E., and Xoplaki, E.: High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects, Holocene, 19, 3–49, 2009. 265, 266, 293

Juckes, M. N., Allen, M. R., Briffa, K. R., Esper, J., Hegerl, G. C., Moberg, A., Osborn, T. J., and Weber, S. L.: Millennial temperature reconstruction intercomparison and evaluation, Clim. Past, 3, 591–609, doi:10.5194/cp-3-591-2007, 2007. 298, 299, 300, 303, 316, 317

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. 265, 290, 291, 302

Krishnamurti, T. N., Kishtawal, C. M., LaRow, T. E., Bachiochi, D. R., Zhang, Z., Williford, C. E., Gadgil, S., and Surendran, S.: Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble, Science, 285, 1548–1550, 1999. 265

Krivova, N. A., Solanki, S. K., Fligge, M., and Unruh, Y. C.: Reconstruction of solar irradiance variations in cycle 23: is solar surface magnetism the cause?, Astron. Astrophys., 399, L1–L4, 2003. 288

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. 288, 289, 291, 292

Mann, M. E., Zhang, Z., Hughes, M. K., Bradley, R. S., Miller, S. K., Rutherford, S., and Ni, F.: Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia, P. Natl. Acad. Sci. USA, 106, 13252–13257, 2008. 298

Mann, M. E., Zhang, Z., Rutherford, S., Bradley, R. S., Hughes, M. K., Shindell, D., Ammann, C., Faluvegi, G., and Fenbiao, N.: Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly, Science, 326, 1256–1260, 2009. 265

CPD
8, 263–320, 2012

Marsland, S. J., Haak, H., Jungclaus, J. H., Latif, M., and Roeske, F.: THe Max Planck Institute global ocean/ice model with orthogonal curvilinear coordinates, Ocean Modell., 5, 91–127, 2003. 291

Moberg, A. and Brattström, G.: Prediction intervals for climate reconstructions with autocorrelated noise – An analysis of ordinary least squares and measurement error methods, Palaeogeogr. Palaeocl., 308, 313–329, 2011. 276

Moberg, A., Sonechkin, D., Holmgren, K., Datsenko, N., and Karlen, W.: Highly variable northern hemisphere temperatures reconstructed from low- and high-resolution proxy data, Nature, 433, 613–617, 2005. 298

Mu, Q., Jackson, C. S., and Stoffa, P. L.: A multivariate empirical-orthogonal-function-based measure of climate model performance, Journal of Geophysical Research, 109, D15101, doi:10.1029/2004JD004584, 2004. 266, 272

Muscheler, R., Joos, F., Beer, J., Müller, S. A., and Vonmoos, M, S. I.: Solar activity during the last 1000 yr inferred from radionuclide records, Quarternary Sci. Rev., 26, 82–97, 2007. 290

Roeckner, E., Bäuml, G., Bonaventura, L., Brokopf, R., Esch, M. G., Hagemann, S., Kirchner, I., Kornblueh, L., Manzini, E., Rhodin, A., Schlese, U., Schulzweida, U., and Tompkins, A.: The atmospheric general circulation model ECHAM5, Part I: Model description, Technical Report, Max Planck Institute of Meteorology, 349, available from MPI for Meteorology, Bundesstr. 53, 20146 Hamburg, Germany, 127 pp., 2003. 291

Rybski, D., Bunde, A., and von Storch, H.: Long-term memory in 1000-year simulated temperature records, J. Geophys. Res., 113, D02106, doi:10.1029/2007JD008568, 2008. 295

Sachs, J. P., Sachse, D., Smittenberg, R. H., Zhang, Z., Battisti, D. S., and Golubic, S.: Southward movement of the Pacific intertropical convergence zone AD 1400–1850, Nat. Geosci., 2, 519–525, 2009. 265

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. 289, 290

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. 289, 290, 295

Shapiro, A. I., Schmutz, W., Rozanov, E., Schoell, M., Haberreiter, M., Shapiro, A. V., and Nyeki, S.: A new approach to the long-term reconstruction of the solar irradiance leads to large historical solar forcing, Astron. Astrophys., 529, 1–8, 2011. 289

Steinhilber, F., Beer, J., and Fröhlich, C.: Total solar irradiance during the Holocene, Geophys. Res. Lett., 36, L19704, doi:10.1029/2009GL040142, 2009. 289, 291

Swingedouw, D., Terray, L., Cassou, C., Voldoire, A., Salas-Mélia, D., and Servonnat, J.: Natural forcing of climate during the last millennium: fingerprint of solar variability, Clim. Dynam., 36, 1349–1364, 2011. 289

Tapping, K. F., Boteler, D., Charbonneau, P., Crouch, A., Manson, A., and Paquette, H.: Solar magnetic activity and total irradiance since the Maunder Minimum, Solar Phys., 246, 309–326, 2009. 289, 291

Wang, Y.-M. and Sheeley, N. R.: Modeling the Sun's Large-Scale Magnetic Field during the Maunder Minimum, Astrophys. J., 591, 1248, 2003. 289

Wenzler, T., Solanki, S. K., Krivova, N. A., and Frohlich, C.: Reconstruction of solar irradiance variations in cycles 21–23 based on surface magnetic fields, Astron. Astrophys., 460, 582–595, 2006. 288

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. 266

Zorita, E., von Storch, H., Gonzalez-Rouco, F. J., Cubasch, U., Luterbacher, J. U., Legutke, S., Fischer-Bruns, I., and Schlese, U.: Climate evolution in the last five centuries simulated by an atmosphere-ocean model: global temperatures, the North Atlantic Oscillation and the Late Maunder minimum, Meteorol. Z., 13, 271–289, 2004. 289

CPD
8, 263–320, 2012

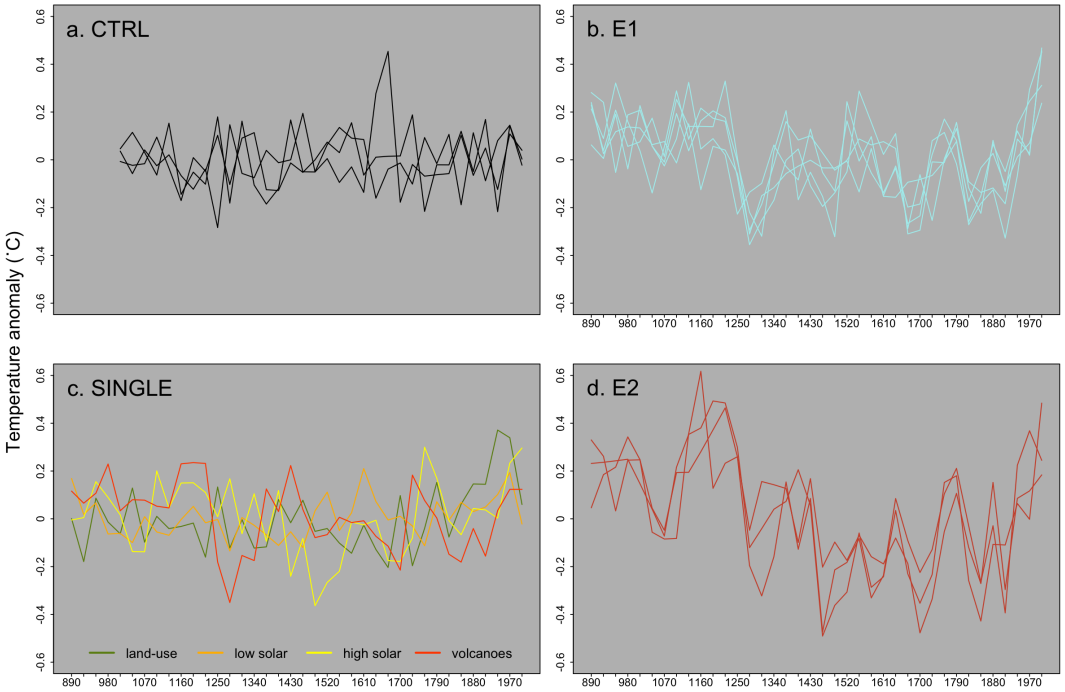

Fig. 1. The MPI Millennium Activity COSMOS simulations over the last millennium with 30-year non-overlapping means of global land-only annual temperature anomalies (°C) from the period 850–2000. The simulations are shown as the CTRLs (top-left panel), E1 ensemble (top-right panel), SINGLE forcing (bottom-left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

CPD
8, 263–320, 2012

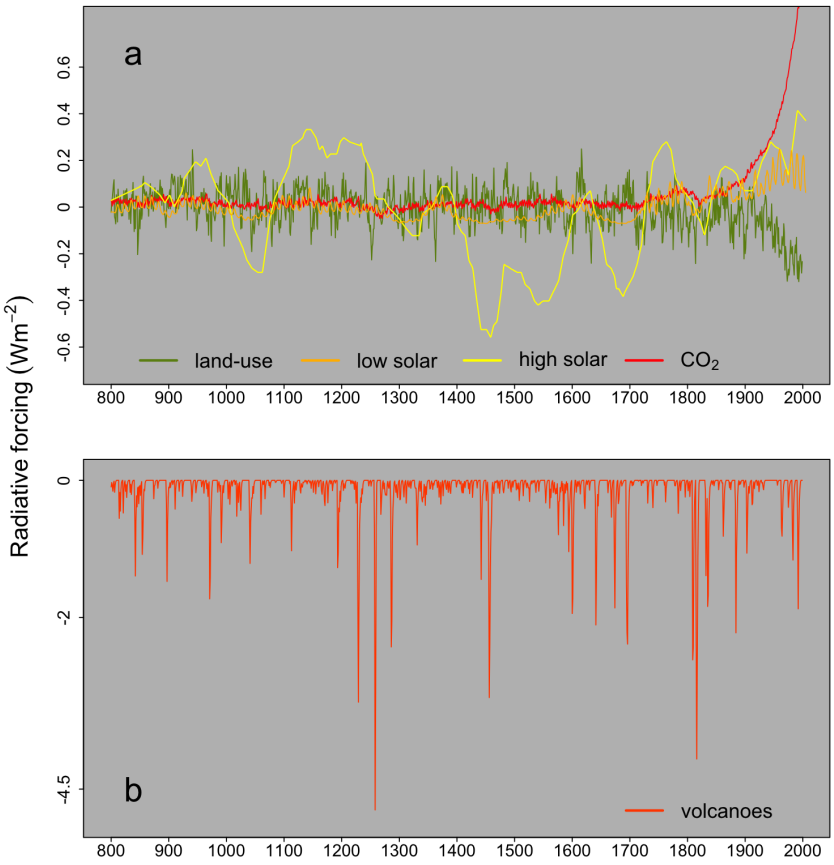

Fig. 2. Annual mean radiative forcing at the top of the atmosphere (Wm^{-2}) for **(a)** low solar (light orange), high solar (yellow), CO_2 (red) and land-cover change (green); and for **(b)** volcanoes (dark orange).

CPD
8, 263–320, 2012

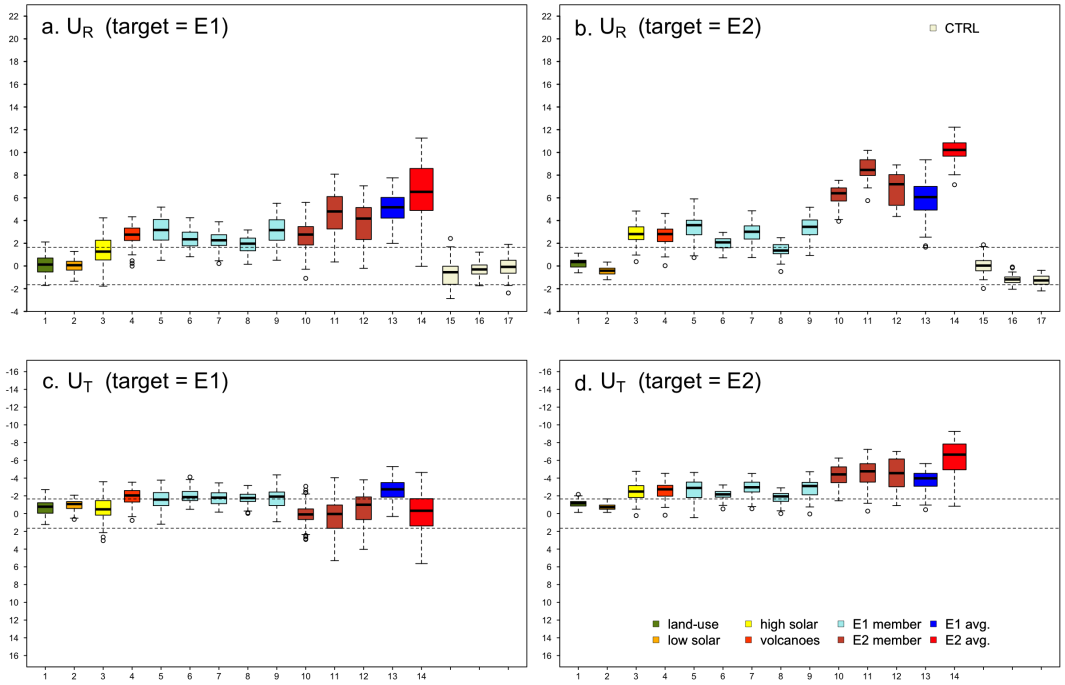

Fig. 3. Box-plots for U_R correlation (top panel) and U_T distance (bottom panel) measures for the global land-only average COSMOS simulation time-series, compared to ≈100 different pseudo-proxy realizations (iteratively running through the ensemble members as targets – *see text*). The left panels are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

CPD
8, 263–320, 2012

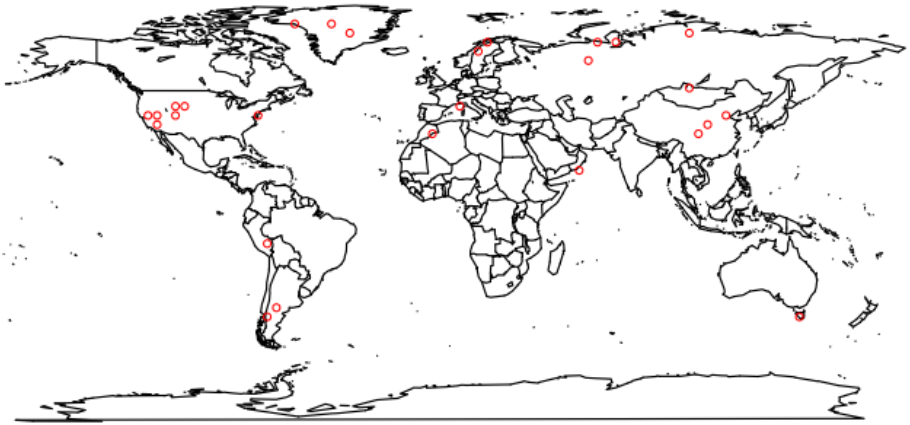

Fig. 4. The 27 proxy locations taken from Juckes et al. (2007) for the present local-scale comparison. Note that the Juckes et al. (2007) set consists of 33 proxy locations, but some locations were so close together a single representation was chosen for that location. A higher resolution model would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

CPD
8, 263–320, 2012

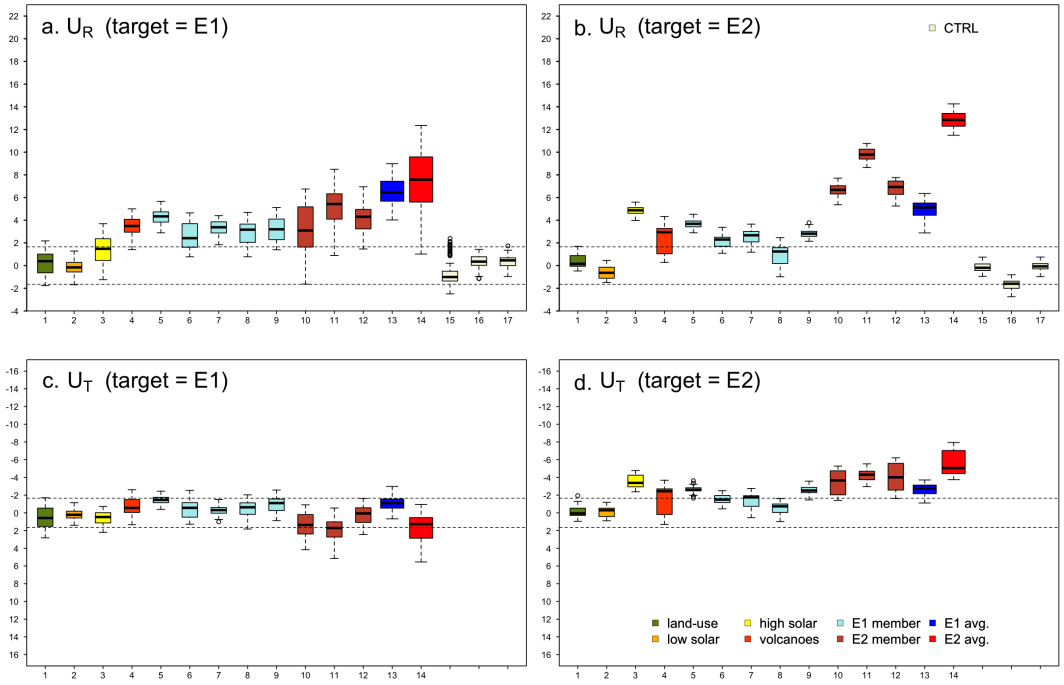

Fig. 5. As Fig. 3, but using the local proxy locations from Juckes et al. (2007).

CPD
8, 263–320, 2012

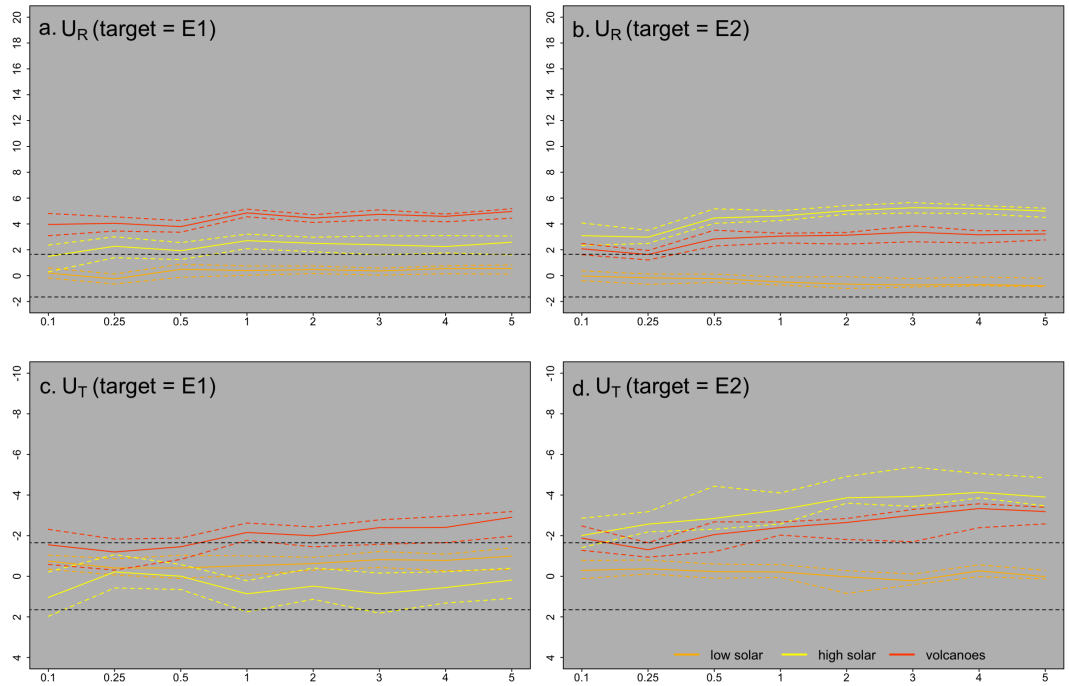

Fig. 6. U_R correlation (top panels) and U_T distance (bottom panels) measures for volcanic (dark orange), low (light orange) and high (yellow) solar forcing simulations. The left panels are for E1 as target, right panels are E2 as target. The 5 % significance level is shown with dashed lines. The filled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

CPD
8, 263–320, 2012

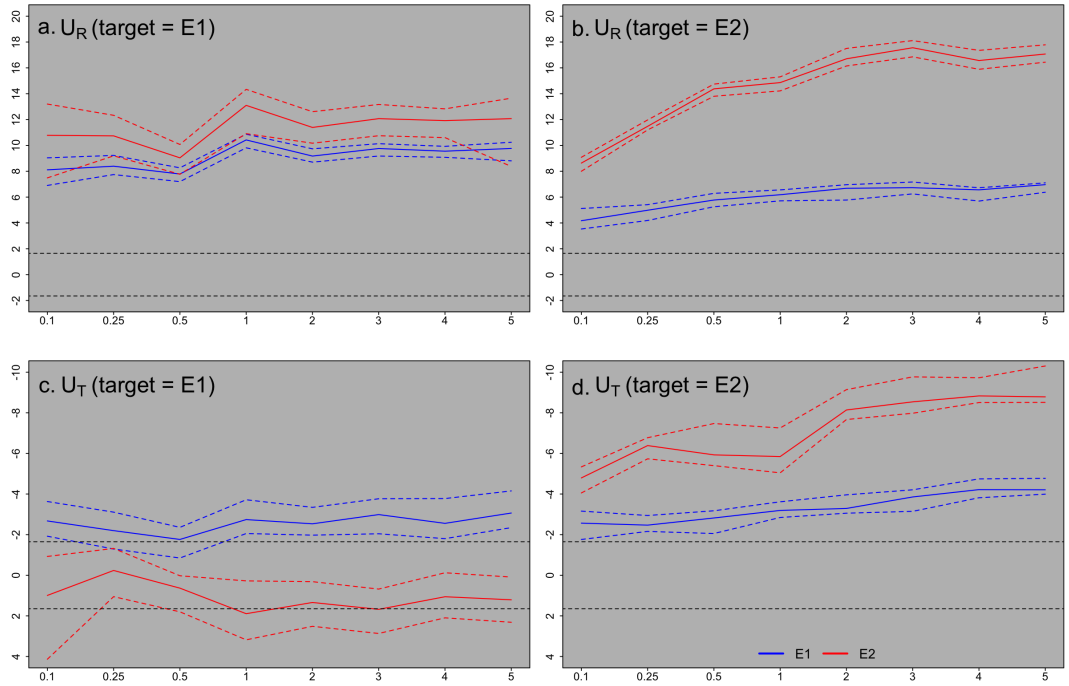

Fig. 7. As Fig. 6, but for the E1 (blue) and E2 (red) ensemble averages.

CPD
8, 263–320, 2012

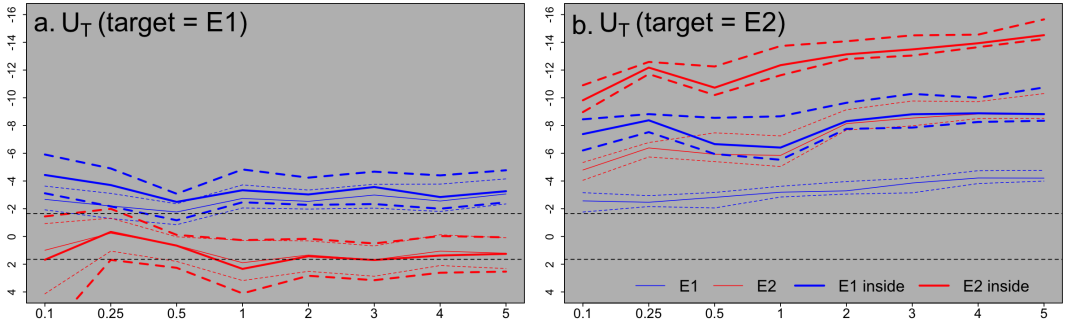

Fig. A1. As Fig. 7, with E1 (blue) and E2 (red) ensemble averages. The thick lines denote the use of inside averaging, whilst the thin lines denote outside averaging (as presented in Fig. 7).

CPD
8, 263–320, 2012