Clim. Past, 8, 1339–1353, 2012
www.clim-past.net/8/1339/2012/
doi:10.5194/cp-8-1339-2012

Climate of the Past

Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium – Part 1: Theory

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

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

Correspondence to: R. Sundberg (rolfs@math.su.se)

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

Abstract. A statistical framework for comparing the output of ensemble simulations from global climate models with networks of climate proxy and instrumental records has been 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 or greenhouse gas concentrations. Two statistical tests have been formulated. Firstly, a preliminary test establishes 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 goal is to objectively rank several competing climate model simulations (e.g. with alternative model parameterizations or alternative forcing histories) by means of their goodness of fit to the unobservable true past climate variations, as estimated from noisy proxy data and instrumental observations.

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, has been made at a regional (European) scale for the last five centuries (Hegerl 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 (Murphy, 1988; Murphy and Epstein, 1989; 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 paleoclimate reconstruction data (Hegerl et al., 2007a), as well as techniques for data assimilation of climate proxy data in model simulations (Goosse et al., 2006; Widmann et al., 2010). Statistical measures of climate model performance can use spatial and temporal correlations found in internal climate variability (Rowlands et al., 2012) 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. For example, we assume that previous model evaluations have shown that the climate models have acceptable biases and that the models, when driven by historical external forcings, simulate climate trends that are consistent with the instrumental observations. 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 far 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.

Our goal here is to develop a method that can be used to objectively rank several "competing" simulations by means of their goodness of fit to the unobservable true past climate variations, as estimated from noisy proxy data and instrumental observations. The competing simulations can be, for example, simulations driven with alternative plausible amplitudes of past radiative forcings (Jungclaus et al., 2010), simulations from models with different climate sensitivities due to different parametrization of unresolved physical processes (Murphy et al., 2004), or a combination of alternative forcings and model parameters is used (Rowlands et al., 2012). We also consider the ranking of entire competing ensembles of simulations, where the members of one and the same ensemble are assumed to differ only in the initial conditions; i.e. their forcings and model physics are the same. Our performance metric will also serve as a test statistic of the null hypothesis that the climate model simulation under consideration is equivalent to an unforced model (control simulation). In addition, we will suggest another test statistic to test the null hypothesis that a climate model simulation does not explain any of the temporal variation in the instrumental or proxy data. To investigate the performance of our framework, under conditions where the results can be evaluated against a perfectly known climate, we undertake a pseudo-proxy experiment in a companion paper (Hind et al., 2012; henceforth Part 2).

We start by formulating a statistical model with 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.

2 A statistical model

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, say, averages over a 10-yr or 30-yr 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. The following notations will be used:

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 as a spatial and temporal average over the same region and 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 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 surrogate for the true temperature τ, derived from climate proxy data. When observed temperatures y are not available, proxy measurements z 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 (x) and in the actual temperature (i.e. all of τ, y and z). This is crucial, since inclusion of temporally varying external forcings in the climate model simulation is the only reason to expect any temporal correlation (covariation) 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, in general, a particular 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 needs 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 latent relationship between the true response to a real forcing and the simulated response to a reconstructed forcing of the same type, when imposed upon a climate model, is (approximately) proportional, when measured as deviations from the mean values of τ and x. This does not prevent additional uncorrelated random forcing effects in the climate model, due to causes discussed above. Note that we make no assumption about how the response to the forcing is related to the forcing itself, but only that the real and simulated responses are linearly related. This way of thinking about the response to climate forcings is similar to that used in detection and attribution studies (e.g. IDAG, 2005). We will discuss, in Sect. 9, some relationships between our framework and that used in detection and attribution.

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 naturally vary. Quantities δ, η, θ, and ϵ are regarded as random variables, with mean values zero and variances σ_δ^2, σ_η^2, σ_θ^2, and σ_ϵ^2, 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 only partially 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 can be linearly explained by the true forcing effect. 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 $\alpha = 1$, whereas an unforced climate model has $\alpha = 0$.

– η denotes the (residual) variation in true temperature that cannot be statistically explained by the particular forcing under consideration. This is then uncorrelated with ξ.

– δ represents internal noise variability in the simulations and any variability in the simulations unrelated (uncorrelated) with the true forcing effects. It will thus incorporate the uncorrelated part of nonlinear climate model effects corresponding to true climate forcing effects.

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

– $\beta(\tau_i - \mu_\tau)$ 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 τ.

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 positively with that in the true temperature τ (or y or z).

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 could be of interest to consider the more complicated case of multiple forcings (i.e. the joint effect of several individual but possibly interacting forcings, not equally represented in the simulations), represented by a vector $\boldsymbol{\xi}$ instead of a scalar ξ. Although climate model simulations driven by multiple forcings are used in the experimental companion paper (Part 2), the theoretical aspects of multiple forcings will be investigated further in a future analysis.

The internal variability sequence δ (but neither θ nor ϵ) will be assumed to be temporally uncorrelated, i.e. white noise, whenever a specification is needed (essentially only for properties of the test statistics of Sects. 6 and 8). This is a slight limitation of the present version of the theory, because the real simulation processes will certainly show some degree of autocorrelation, at least on rather short time scales. In the pseudo-proxy experiment in Part 2, white noise δ is assumed, motivated by finding and selecting a time unit for which the autocorrelation is negligible.

We cannot expect a forced simulation to explain some of the variability in the real temperature, unless it is statistically correlated. Thus, in practice, if we want to test some forced simulations of different types, to rank them according to how well they are able to explain the observed temperatures, it is natural to first test by correlation tests whether they can explain any of the observed temperature variations. Only forcings that provide statistically significant correlations between simulated and observed temperatures 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 carried out 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:

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

A statistical motivation for this criterion is obtained by considering D^2 as a mean squared error of prediction (MSEP) of τ.

It may be argued that such an MSEP criterion function should be formulated as dependent on possible autocorrelation among the simulation errors δ_i in x_i (Mu et al., 2004). If we assume δ_i to be white noise, this argument disappears. Even without this assumption, we can choose criterion (1), primarily for its simplicity. We will return to motivations in Sect. 5. The requirement on δ_i to be white noise will reappear in the calculation of standard errors for the test statistics, but even there, this assumption is not crucial so long as we use these statistics more as criteria for ranking than for stating significances.

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 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:

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

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 will yield a procedure for the calibration of the proxy series z, 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 Proxy calibration to avoid ranking bias in $D^2(x, z)$

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 ranking in any systematic way. More precisely, we demand the mean value of $D^2(x, z) - D^2(x, \tau)$, given x and τ, to be free of x and τ, in particular free of $x - \tau$. We first conclude that replacement of τ by y in $D^2(x, \tau)$ does not introduce any ranking bias. This is seen from the relation

$$(x - y)^2 - (x - \tau)^2 = -2(x - \tau)\theta + \theta^2. \quad (3)$$

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

For the proxy data, following the statistical model, the formula corresponding to Eq. (3) is

$$(x - z)^2 - (x - \tau)^2 = -2(x - \tau)(z - \tau) + (z - \tau)^2, \quad (4)$$

where

$$z - \tau = \mu_z - \mu_\tau + (\beta - 1)(\tau - \mu_\tau) + \epsilon. \quad (5)$$

In the mean value of Eq. (4), the contribution of the noise term ϵ will only be the constant σ_ϵ^2, corresponding to σ_θ^2 from Eq. (3). For the mean value of Eq. (4), given x and τ, to be totally free of $x - \tau$, however, it is seen from Eq. (5) that we must have both $\mu_z = \mu_\tau (= \mu_y)$ and $\beta = 1$. This tells how raw proxy data z must be calibrated, so that it does not introduce any systematic deviations from the ideal ranking.

For the calibration of the proxy data z, we assume that we have available a period of both proxy and temperature measurements. This allows the estimation of the relationship between z and y, and this is the basis for the calibration. The first requirement, that z should have the same mean value as y, is naturally achieved except for the unavoidable calibration uncertainty by adjusting z by an additive constant, so its average value $\overline{z}$ over the calibration period satisfies $\overline{z} = \overline{y}$.

The second requirement is that z should have a regression on the latent variable τ with regression coefficient $\beta = 1$. Thus, given an uncalibrated proxy z_0 with regression coefficient β_0 on τ, z_0 should be rescaled by division with β_0 to form a calibrated z as $z = \mu_y + (z_0 - \mu_{z_0})/\beta_0$, except for calibration uncertainty. In the case when the error in y is negligible, so $y = \tau$, this corresponds precisely to the so-called classical calibration procedure (Osborne, 1991; 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), we can obtain an approximately unbiased estimator of β_0 by

$$\widehat{\beta}_0 = \frac{s_{yz_0}}{s_y^2 - \widehat{\sigma}_\theta^2}, \quad (6)$$

where s_{yz_0} and s_y^2 are the empirical covariance and variance, respectively (see Fuller, 1987; Cheng and van Ness, 1999). 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, when errors in y are negligible. To allow errors in y, the modified Eq. (6) should be used.

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. (6). The reason is that the full amplitude of the true temperature signal is retained and that the proxy noise variance is superimposed on the temperature signal variance. We will see in the next section how this is compensated for by an optimal weighting of the different observed values according to their variance components.

Remark. A calibrated proxy z, obtained by classical calibration or by using Eq. (6), may be called an estimator or predictor of τ in the sense that the true temperature component embedded within the noisy proxy series is estimated with its correct variance. However, the weaker the correlation is between z and τ (or y), the larger the total variance in z because the variance of the noise term ϵ becomes increasingly dominant. The z calibrated in this way is therefore not an optimal predictor of the true temperatures at each individual time point. For a single time-point prediction, direct regression of τ (or y) on z_0 would provide a more appropriate predictor/estimator, where the prediction error variance is minimized. This alternative way of calibrating climate proxy data has often been used in palaeoclimate studies (NRC, 2006). For the climate reconstruction problem in general, though, the seemingly desirable property of (in theory) minimized prediction error is not necessarily an advantage, because (in practice) it leads to a systematic bias of the mean reconstructed (i.e. predicted) past climate in periods that have a mean value that differs from that of the calibration period. As this is an undesired property, e.g. when judgements are made on how the recent climate differs from previous climates, this has led to vigorous discussions in the climate literature on how proxy data should be calibrated. von Storch (1999); Esper et al. (2005a); Bürger et al.

(2006); Hegerl et al. (2007a); Ammann et al. (2010); Christiansen (2011); Kutzbach et al. (2011); Moberg and Brattström (2011); Tingley et al. (2012) and several others have recently discussed the importance of retrieving the full variance of the temperature signal, including discussions on the use of errors-in-variables models. It must be stressed, however, that in the present context, the way proxy data should be calibrated comes out as a corollary from the statistical model formulated in combination with the explicit desire to obtain an unbiased ranking of forced simulations against the true past temperatures.

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, a number of complications must be dealt with, e.g. different time periods for different proxies, and there are several reasons (e.g. collinearity among the proxies, or that the relationships from the calibration period do not hold outside this period) why another way to combine the proxies may be preferred. We will not attempt to deal with these more practical problems here, but primarily 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. (6). In cases when different proxy data are available in different pre-instrumental periods, they must be separately calibrated, and when the calibrated proxy series is known to have different precision in different time periods, this must be adjusted for in the weighting (see Sect. 5 below).

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. 10- or 30-yr 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. However, this will of course decrease the statistical precision. The instrumental noise variance to be used in Eq. (6) can be difficult to estimate in practice, but efforts to estimate errors in gridded temperature data have been made (Brohan et al., 2006). Moreover, Moberg and Brattström (2011, Sect. 6.1) discuss a procedure to estimate the error variance in the mean of a set of neighbouring temperature station records.

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

Direct temperature measurements y and proxies z have 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 this weighting in D^2, we first reconsider Statistical Model 1, assuming both $\alpha = 1$ (correct amplitude of the forced component 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 $x - z$.

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

$$e^{-\frac{n}{2} D^2(x,z)/(\sigma_\delta^2+\sigma_\eta^2+\sigma_\epsilon^2)}, \tag{7}$$

where σ_δ^2, σ_η^2 and σ_ϵ^2 are the variances of the corresponding components of the statistical model, and $\mathrm{Var}(x_i - z_i) = \sigma_\delta^2 + \sigma_\eta^2 + \sigma_\epsilon^2$. 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 (7) to observe this vector $x - z$ will tend to be exponentially smaller. Thus, the D^2 measure is proportionally equivalent to a Gaussian log-density.

The denominator $\sigma_\delta^2 + \sigma_\eta^2 + \sigma_\epsilon^2$ in the exponent of Eq. (7) was assumed constant. However, when the variances in this denominator vary with i, in particular the proxy noise term $\sigma_\epsilon^2(i)$, the interpretation of D^2 as a Gaussian log-density tells us how different terms should be (ideally) weighted in D^2, forming a weighted version D_w^2:

$$D_\mathrm{w}^2(x,\, z) = \frac{1}{n}\sum_1^n w_i\,(x_i - z_i)^2 = \frac{1}{n}\sum_1^n \frac{(x_i - z_i)^2}{\sigma_\delta^2 + \sigma_\eta^2 + \sigma_\epsilon^2(i)}. \tag{8}$$

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. (7). We will use that version as our definition for w_i:

$$w_i = \frac{\sigma_\delta^2 + \sigma_\eta^2}{\sigma_\delta^2 + \sigma_\eta^2 + \sigma_\epsilon^2(i)}. \tag{9}$$

For times i when a precise $y = \tau$ is available (i.e. with $\sigma_\epsilon^2(i) = \sigma_\theta^2 = 0$), the normalized weight (Eq. 9) equals 1, whereas $w_i < 1$ when a noisy proxy z is used, or an imprecise instrumental y.

Since the denominator of Eq. (8) is the expected value of the numerator of the same term, an alternative interpretation of the proposed weights is that they are chosen to make all terms of D_w^2 be of the same magnitude.

The weight factor introduced in Eq. (9) is an ideal weight (under the assumptions made), for which we can at best give an estimate. Thus, we must insert estimates for each of the three variance 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 have reason to allow $\sigma_\epsilon^2(i)$, and thus also the weight w_i, to vary

over time depending on the precision of the available instrumental or proxy measurement.

To estimate σ_δ^2, we propose to use the (time-)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 unknown forcing effect 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 to be ranked.

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. (9), 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 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 $\sigma_\epsilon^2(i)$ in Eq. (9) 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:

$$w_i = \frac{s_\delta^2 + s_y^2}{s_\delta^2 + s_y^2/\rho^2(y, z)} \tag{10}$$

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 is 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 are used becomes

$$w_i = \frac{s_\delta^2 + s_y^2(1-q)}{s_\delta^2 + s_y^2}. \tag{11}$$

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

$$w_i = \frac{s_\delta^2 + s_y^2(1-q)}{s_\delta^2 + s_y^2(1-q)^2/\rho^2(y,\, z)}. \tag{12}$$

A minor drawback (in *practice*) of Eq. (12) is that it might generate weights $w_i > 1$. This occurs when $\rho^2 > 1 - q$ for the estimated values of ρ and q (not being possible for the *true* values). Should that happen, we advise that the estimation procedures for q and ρ are checked, and the assumptions of uncorrelated noise in, and between, θ and ϵ. As a resort, if this does not help, w_i could be redefined by using Eq. (10) or (11), bearing in mind that the resulting weights are not optimal.

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

When a D^2 value is 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 this question, we construct a statistical test of a null hypothesis expressing that the forced model is not better than an unforced model:

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

Since the unforced model (control simulation) is important here as a reference, it will be given a specification separate from the general Statistical Model 1 in Sect. 2.

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

$$\begin{aligned} x_i &= \mu_x + \delta_i \\ \tau_i &= \mu_\tau + \eta_i \\ y_i &= \tau_i + \theta_i \\ z_i &= \mu_z + \beta\,(\tau_i - \mu_\tau) + \epsilon_i \end{aligned}$$

where δ_i (but still neither η_i, θ_i nor ϵ_i) is regarded as white noise, for the variance formula below.

Note that the previous forced component of τ, i.e. ξ in Statistical Model 1, is now included in η, because there is no

longer a point in expressing it explicitly. This means the *true* temperature τ can include forcing effects also in Model 2, albeit only implicitly. Except for that change of interpretation of η, the observed climate part of the model is the same as in Model 1. For the simulation variance σ_δ^2, there is no clear general answer. If the incorporation of a forcing effect in the climate model does not increase the overall variance in the simulations, the residual variance σ_δ^2 must shrink. Another possibility is that the forcing effect is simply added to the variation and thus σ_δ^2 is not affected. However, to the extent the forcing effects in the climate model are not precisely proportional to those in the true climate, they will contribute to an increased σ_δ^2, so this is also a possible scenario.

We will not deal further with this problem here, but, when necessary, simply assume that σ_δ^2 is the same both with and without forcings. A somewhat related approach to the problem of comparing climate models with the same types of forcings, but with different magnitudes, would be to try estimating α. Again, this will be a topic for future study.

The unforced climate 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:

$$T\,(x_\mathrm{f},\, x_\mathrm{u},\, z) = \overline{D^2_\mathrm{w}}\,(x_\mathrm{f},\, z) - \overline{D^2_\mathrm{w}}\,(x_\mathrm{u},\, z) \tag{13}$$

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 $T\,(\overline{x}_\mathrm{f},\, \overline{x}_\mathrm{u},\, z) = D^2_\mathrm{w}(\overline{x}_\mathrm{f},\, z) - D^2_\mathrm{w}(\overline{x}_\mathrm{u},\, z)$.

This alternative procedure would be even more efficient but is not used here, because it would also introduce a bias in the comparison, unless $k = K$. However, we provide details necessary to use this alternative in Appendix A and both variants are used in our experiments in Part 2.

We show below that an approximate distribution under H_0 for the test statistic in Eq. (13) can be obtained with the help of an analytical formula for its standard error. In doing this, we will regard the z (and y) series as fixed and given. It means that we do not need any distributional assumptions about the z series. This is possible because z is common to both terms of Eq. (13).

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, which will be denoted μ_x. The test statistic value can be rewritten as

$$\begin{aligned} T\,(x_\mathrm{f},\, x_\mathrm{u},\, z) = {} & \overline{w\,(x_\mathrm{f} - \mu_x)^2} - \overline{w\,(x_\mathrm{u} - \mu_x)^2} \\ & - 2\,\overline{w\,(\overline{x}_\mathrm{f} - \overline{x}_\mathrm{u})\,(z - \mu_x)}, \end{aligned} \tag{14}$$

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. (14) are sums of a large number of terms (referring to the central limit theorem of probability). Under H_0, the expected value of $T\,(x_\mathrm{f},\, x_\mathrm{u},\, z)$ is zero, since the forced climate model is equivalent to 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

$$\begin{aligned} \mathrm{Var}(T\,(x_\mathrm{f},\, x_\mathrm{u},\, z)) = {} & \frac{1}{n^2}\left(\frac{1}{k} + \frac{1}{K}\right) \\ & \left\{2\sigma_\delta^4 \sum_1^n w_i^2 + 4\sigma_\delta^2 \sum_1^n w_i^2\,(z_i - \mu_x)^2\right\}. \end{aligned} \tag{15}$$

An approximately $N(0, 1)$-distributed test statistic is obtained by normalizing the T-value in question by its standard error, i.e. by the square root of Eq. (15) 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 negative, 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 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. 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. (15). This is not discussed further here (but 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* (statistical precision) 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:

$$\mathrm{Var}\left(\sum_j T_j\right) = \sum_j \mathrm{Var}\left(T_j\right) + 2 \sum_{j_1<j_2} \mathrm{Cov}\left(T_{j_1}, T_{j_2}\right). \quad (16)$$

Thus, what we need, together with the variances discussed in the previous section, are the covariances. Consequently, we need to supplement the variance Eq. (15), by the corresponding formula for covariances. This is obtained for $\mathrm{Cov}(T_{j_1}, T_{j_2})$ from Eq. (15) by the following three operations:

- Change σ_δ^2 to $\mathrm{Cov}(\delta(j_1), \delta(j_2))$, and σ_δ^4 to the covariance squared.
- Change w_i^2 to $w_i(j_1)\, w_i(j_2)$.
- Change $(z_i - \mu_x)^2$ to

  $$(z_i(j_1) - \mu_x(j_1))\,(z_i(j_2) - \mu_x(j_2)).$$

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 here assumed that not only are the variances σ_δ^2 constant over time, as in Eq. (15), but also the corresponding covariances. Note that the first term in the sum contains a covariance squared, corresponding to s_δ^4 in the variance Eq. (15). 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. (15).

Now let nonequal weights be allowed, in the form $\sum c_j\, T_j$, 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 $\boldsymbol{c}$ be the corresponding column vector with components c_j. Then the variance for $\sum_j c_j\, T_j$ is obtained as the scalar:

$$\mathrm{Var}\left(\sum_j c_j\, T_j\right) = \mathbf{c}^T\, \mathbf{V}(\boldsymbol{T})\, \boldsymbol{c}. \quad (17)$$

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

$$U_T = \frac{\sum_j c_j\, T_j}{\sqrt{\mathrm{Var}\left(\sum_j c_j\, T_j\right)}}. \quad (18)$$

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 the U_T-value is, the better (note that a forced model with $U_T > 0$ performs worse than an unforced model).

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 $\mathrm{Var}(T)$ and $\mathrm{Cov}(T)$ 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. For a particular shorter proxy series, we simply let $w_i = 0$ for all time points i, where we have no measurement. A consequence of this is that more weight will be given to regional/seasonal data with long proxy series than those with short series, which seems reasonable. Note, as an example, that a mean value μ_x in Eq. (15) for each region should be interpreted as only representing the time period when proxy data are available for that region, so the actual $\overline{z}$ can be a natural estimate of μ_x.

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 include 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, 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 Part 2, we will simply use the same weight for all regions.

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), but (positively) correlated when forcing effects appear jointly in 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.*

Under H_0, we should of course not expect any significant correlation or covariance between x and z. 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:

$$R(x, z) = \frac{\sum \widetilde{w}_i \, (\overline{x}_i - \mu_x) \, (z_i - \mu_z)}{\sum \widetilde{w}_i^2 \, (z_i - \mu_z)^2} \tag{19}$$

for a given z series. We allow k 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, i.e. the square root of its variance,

$$\mathrm{Var}(R(x, z)) = \frac{(1/k) \, \sigma_\delta^2}{\sum \widetilde{w}_i^2 \, (z_i - \mu_z)^2}, \tag{20}$$

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, $\overline{z} = \sum \widetilde{w}_i \, z_i / \sum \widetilde{w}_i$.

The weight factor $\widetilde{w}_i$, however, is *not* the same weight factor w_i 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 no, or negligible, noise are used): $\widetilde{w} = 1$.
2. If $y = \tau + \theta$ (instrumental data with non-negligible noise variance, variance proportion q): $\widetilde{w} = 1 - q$.
3. If $y = \tau$, $z = \tau + \epsilon$ (proxy data are used, no noise in y, calibration period yields $\rho^2(y, z)$): $\widetilde{w} = \rho^2(y, z)$.
4. If $y = \tau + \theta$, $z = \tau + \epsilon$ (proxy data are used, noise in y, calibration period yields $\rho^2(y, z)$): $\widetilde{w} = \rho^2(y, z)/(1 - q)$.

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 can expect that adequate forcing effects in the simulations will covary with the actual variation in the z sequence and therefore contribute to a significant test outcome.

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 following formula:

$$\mathrm{Cov}\left(R_{j_1}, R_{j_2}\right) = \frac{(1/k) \, \rho\left(\delta_1, \delta_2\right) \, \sigma_{\delta_1} \, \sigma_{\delta_2} \sum \widetilde{w}_{1i} \, \widetilde{w}_{2i} \left(z_{1i} - \mu_{z_1}\right) \left(z_{2i} - \mu_{z_2}\right)}{\sum \widetilde{w}_{1i}^2 \left(z_{1i} - \mu_{z_1}\right)^2 \sum \widetilde{w}_{2i}^2 \left(z_{2i} - \mu_{z_2}\right)^2}. \tag{21}$$

Here ρ is the coefficient of correlation between two joint x-sequences from a single simulation, to be estimated from a set of (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 $\boldsymbol{c}$ be the corresponding column vector with components c_j. Then the variance for $\sum_j c_j \, R_j$ is obtained as the scalar:

$$\mathrm{Var}\left(\sum_j c_j \, R_j\right) = \boldsymbol{c}^T \, \mathbf{V}(\boldsymbol{R}) \, \boldsymbol{c}. \tag{22}$$

Thus, our unified correlation-based test statistic becomes

$$U_R = \frac{\sum_j c_j \, R_j}{\sqrt{\mathrm{Var}\left(\sum_j c_j \, R_j\right)}}. \tag{23}$$

A significant positive value of U_R implies that the model is able to explain some of the observed temperature variation. 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 one-sided 5 % significance level. The larger the positive U_R values are, the stronger the correlation between the model and the observations. A high negative U_R would imply a negative correlation between model and observations. If such values are found, this may be a warning sign 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-based significance test of R (see Appendix B).

9 Discussion – relationships between our framework and optimal fingerprinting used in detection and attribution studies

As pointed out by Hegerl (2012), the statistical framework developed here has similarities to the ideas underlying the optimal fingerprinting method used in detection and attribution (DA) studies, which have been important for our understanding of the relative roles of man-made and natural influences on recent climate change (e.g. IDAG, 2005; Hegerl et al., 2007b). Given the importance of the optimal fingerprinting and DA framework, it seems worthwhile discussing some differences and similarities to our framework.

Our framework is specifically developed to obtain an unbiased ranking of several competing forced simulations where a variety of climate proxy data, representing different regions, covering different time periods and having different precision, are used to represent the real climate. The optimal fingerprinting method, although it has indeed been used for comparisons of simulated climate with climate reconstructed from proxy data (e.g. Hegerl et al., 2007a), has largely been designed for DA studies using spatially more complete and homogeneous gridded instrumental climate data sets (e.g. Stott et al., 2003). Initially, in our work, we considered the option of modifying an existing empirical-orthogonal-function-based quadratic form type metric of climate model performance (Mu et al., 2004; Rowlands et al., 2012). This incorporated essential elements from optimal fingerprinting, but it was not clear how to modify it in order to reach our goals, for example allowing noise in simulations, instrumental observations, and proxies – in particular allowing proxy series of different length and with temporally variable statistical precision. Therefore, we preferred to start from scratch and develop a model-based statistical framework specifically designed for our purposes.

A central assumption behind the use of optimal fingerprinting in DA studies is that the observed climate record, which is influenced by multiple sources of external forcings, can be expressed as the sum of internal unforced climate variations plus a linear combination of simulated response patterns to each of the individual forcings as determined from climate model simulations. Thus, the method relies on the existence of an underlying linear relationship between the real response to a real forcing and the expected simulated response to the same type of forcing within a climate model.

Mathematically, but not statistically, the same idea underlies the specification of the simulated (x) and real (τ) temperatures in our Statistical Model 1. Here, the real forced temperature response is represented by the term ξ and the corresponding simulated response is represented by $\alpha\xi$, where α can be interpreted as a linear scaling factor, in a theoretical linear regression of the climate model output on the corresponding true climate forcing effect. Thus, α plays an inverse type of role to the regression coefficients used in optimal fingerprinting to scale the simulated response patterns (signals) such that they best fit the observed climate. In both cases, i.e. in our framework and optimal fingerprinting, a perfectly simulated amplitude of the response to a particular forcing means that the scaling factor should be equal to one. A difference between the two approaches, so far, is that optimal fingerprinting allows a vector of scaling factors to deal with jointly linear effects of different forcings, whereas our framework has been deliberately restricted to a single factor and a single type of forcing (although this may consist of a combination of several individual forcings). However, as mentioned in Sect. 2, we intend to investigate the implications of an extension of our framework to allow explicit treatment of several individual forcings. Under additional assumptions, our framework can in principle be extended to include a regression-type estimation of α, which would correspond to the estimation part of optimal fingerprinting.

The first requirement in any DA study is to determine whether an observed climate change can be detected beyond the level of internal unforced variability (IDAG, 2005). This occurs when the estimated signal pattern scaling factors are significantly different from zero. In our framework, this corresponds to determining whether a forced model is able to explain any of the observed temperature variations. This is achieved by our correlation test statistic U_R defined in Sect. 8. The second DA requirement is to assess the consistency between the observed and simulated response to forcing (IDAG, 2005). This is the same as evaluating the null hypothesis that all fingerprint regression coefficients are equal to one. This part of the DA framework has, so far, no counterpart in our framework. However, after a future modification to allow an estimation of α (also in vector form), our framework could handle this aspect too. Our main test statistic U_T, defined in Sect. 7, has no direct counterpart in optimal fingerprinting, but plays a similar role as a performance metric, namely as the "cost function" defined in Eq. (1) of Mu et al. (2004), or the goodness-of-fit statistic used in Rowlands et al. (2012). Our U_T, however, is not merely a distance metric, but is in fact a test statistic derived from a set of distance metrics calculated for different regions and seasons.

Both our framework and optimal fingerprinting require some assumptions about the spatiotemporal character of the unforced internal climate variability. Optimal fingerprinting typically relies on a spatiotemporal covariance matrix estimated from long control simulations (Allen and Tett, 1999). This matrix can be very large, which complicates its inversion – something that is needed in optimal fingerprinting. This inversion may require non-trivial "pre-whitening" operations (Allen and Tett, 1999). Our framework does not require any inversion of covariance matrices and is therefore easier to use. The price paid for avoiding the need to allow temporal correlation is a simplifying assumption of white noise in the specification of the temporal character of the unforced *simulated* climate (δ in Statistical Models 1 and 2). This limitation places a restriction on the time unit (time resolution) that can be used; namely, this unit must be sufficiently long such that the unforced *simulated* climate can be approximated as white noise. Optimal fingerprinting is in theory not hampered by such a restriction, but in practice it may not be possible to use a much smaller time unit, because that would significantly increase the size of the covariance matrix to be inverted. Moreover, to obtain information about the full spatiotemporal covariance on all timescales relevant for studies of climate of an entire millennium, several very long control simulations would be needed. Thus, it seems questionable whether it is possible, in practice, to take full advantage of the inclusion of the covariance matrix in the optimal fingerprinting framework for multivariate model evaluations that consider the entire last millennium, or longer. It is also debatable whether or not the spatial correlation should go into the criterion in an automatic way (via EOFs), to be used for evaluating climate models. It should, however, certainly go into the evaluation of the properties of the criterion.

Note also that – although not being an explicit part of our statistical framework – we suggest the white noise assumption for δ of the control simulations can be satisfied to a sufficient degree (i.e. not severely violated) by a suitable choice of time unit (see Part 2, where this is investigated). Even better could be to have a physically more realistic character of δ in the model and find the corresponding adjustments required in the theoretical results. Future work could be aimed at this problem. Concerning spatial variation, note that explicit information from control simulations is already included in our framework, for the estimation of variances and covariances needed in the test statistics U_T and U_R. This part of our framework does not put any theoretical restriction on the character of the spatial covariances. We also remind the reader that no assumptions at all are needed for the temporal character of the *real* unforced temperature (η) or the forced temperature component (ξ), nor of the measurement error components θ and ϵ.

10 Conclusions and recommendations

We have presented statistical models for observed and latent variables that play a role when comparing forced climate model simulations with climate proxy data or instrumental records. Based on these models, a distance measure between simulations and proxies has been developed that ranks the simulations in the same order as if the distance to the true, unknown, climate were used – of course with an unavoidable stochasticity due to the noise in the data. This distance measure can be used with a set of multiple proxies that represent different regions and seasons, and includes weights that depend on the statistical precision of these proxies, which is allowed to vary in time. A significance test is then developed, to test if a forced simulation performs better (i.e. has a smaller distance to the observations) than unforced simulations. Another significance test is formulated for the correlation between a forced simulation and the proxies. Although distance measures are a standard concept, this is – to our knowledge – the first time the specific form of the distance measure and the calibration of proxies are jointly developed based directly on a statistical model for comparing simulated and observed past climate records, rather than being ad hoc.

The new framework may be used to rank a set of alternative simulations, where the models are driven with different amplitudes of past external forcings, e.g. solar and volcanic forcing. This may help to better understand how large these past forcings have been, something that is not yet fully understood, by assuming that those reconstructed forcings that provide the best fit of simulated temperatures to the observed ones are more likely to better represent the true past forcings. In our companion study (Part 2) we will investigate, in a pseudo-proxy experiment, the possibility to distinguish between multiple-forced simulations that include either a small or a large amplitude of past solar forcing.

Alternatively, our framework could be used to rank different simulations that have been driven with the same past forcing history, but where the climate models include different parametrizations of various non-resolved physical processes, i.e. a perturbed physics ensemble simulation experiment. Different parametrizations may cause the models to have different climate sensitivities, resulting in different amplitudes of the response to external forcings. The models that provide the best fit to the observations would likely include the more appropriate parametrizations.

Note that our framework is designed for being applied to fully coupled general circulation models (GCMs). Thus, we do *not* advise to rank simulations with a simple energy balance model (EBM), or an Earth system model of intermediate complexity (EMIC), against simulations with a GCM. Such rankings would be rather meaningless. For example, an EBM might provide a smaller distance to the observations than a GCM just because its unforced variability is virtually zero. Also, note that our correlation-based test statistic should *not* be used to rank simulations; its sole purpose is to determine

whether a forced model is able to explain (in a statistical sense) some part of the observed temperature variations. A ranking using the distance-based statistic is meaningful only when this occurs.

The framework developed here is not yet fully developed and several aspects could be improved in future versions. It would be desirable to have a statistical precision measure attached to the D^2 differences when comparing differently forced climate models, not only as now, when forced simulations are only compared with unforced controls. The assumption that the simulated unforced variability can be modelled as white noise should be replaced by a physically more realistic representation, making it possible to work with shorter time units. Also, further modification would be required for use with climate variables other than temperature – or a combination of several climate variables. Future developments should aim at improving some of the aspects mentioned above, as well as to allow estimation of the forced amplitude of simulated climate variability.

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 $T(\overline{x}_\mathrm{f}, \overline{x}_\mathrm{u}, z) = D^2_\mathrm{w}(\overline{x}_\mathrm{f}, z) - D^2_\mathrm{w}(\overline{x}_\mathrm{u}, z)$ 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

$$T(x_\mathrm{f}, x_\mathrm{u}, z) = \overline{D^2_\mathrm{w}}(x_\mathrm{f}, z) - \overline{D^2_\mathrm{w}}(x_\mathrm{u}, z),$$

i.e. 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

$$T(\overline{x}_\mathrm{f}, \overline{x}_\mathrm{u}, z) = D^2_\mathrm{w}(\overline{x}_\mathrm{f}, z) - D^2_\mathrm{w}(\overline{x}_\mathrm{u}, z),$$

i.e. under inside averaging, the expected value of T contains an additional bias term, and is now

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

The bias term is now 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.

A2 Precision of the test statistic T

Under inside averaging, the analytical formula for the variance of T, under assumed normality of $x_\mathrm{f} - x_\mathrm{u}$ and given the z sequence, is

$$\mathrm{Var}\,(T(\overline{x}_\mathrm{f}, \overline{x}_\mathrm{u}, z)) = \frac{2\sigma^4_\delta}{n^2}\left(\frac{1}{k^2} + \frac{1}{K^2}\right)\sum_{1}^{n} w_i^2 + \frac{4\sigma^2_\delta}{n^2}\left(\frac{1}{k} + \frac{1}{K}\right)\sum_{1}^{n} w_i^2\,(z_i - \mu_x)^2.$$

Appendix B

Alternative reference distributions for D^2_w, T and R

It deserves mention that there are (at least) two possible non-parametric alternatives to the normality-based tests for H_0 defined in Sect. 6, 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 them represent x_f under H_0. Along these lines, a reference distribution for the distance measure D^2_w 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 more affordable 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. 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^2_w and T values can be computed, and this leads to a reference distribution for D^2_w 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. However, it should not be forgotten that the primary use of D^2_w is for ranking different simulations using data from several regions and seasons, 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.

Acknowledgements. This research was funded by the Swedish Research Council (grants 70454201, 90751501 and B0334901) and the European Union (FP6 grant 017008, "Millennium" project). We thank G. Hegerl, Y. H. Yamazaki and an anonymous reviewer for constructive comments and advice in their reviews of the discussion paper.

Edited by: P. Brohan

References

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

Ammann, C. M., Genton, M. G., and Li, B.: Technical Note: Correcting for signal attenuation from noisy proxy data in climate reconstructions, Clim. Past, 6, 273–279, doi:10.5194/cp-6-273-2010, 2010.

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

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

Bürger, G., Fast, I., and Cubasch, U.: Climate reconstruction by regression – 32 variations on a theme, Tellus A, 58, 227–235, 2006.

Cheng, C. -L. and van Ness, J. W.: Statistical regression with measurement error, Arnold Publishers, London, UK, 1999.

Christiansen, B.: Reconstructing the NH mean temperature: can uncerestimation of trends and variability be avoided?, J. Climate, 24, 674–692, 2011.

Esper, J., Frank, D. C., Wilson, R. J. S., and Briffa, K. R: Effect of scaling and regression on reconstructed temperature amplitude for the past millennium, Geophys. Res. Lett., 32, L07711, doi:10.1016/j.quascirev.2005.07.001, 2005a.

Fuller, W. A.: Measurement error models, Wiley, New Jersey, USA, 1987.

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.

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.

Hegerl, G. C.: Interactive comment on "Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium" by A. Hind et al., Clim. Past Discuss., 8, C114–C117, 2012.

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.

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.

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.

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

IDAG – The International Ad Hoc Detection and Attribution Group: Detecting and Attributing External Influences on the Climate System: A Review of Recent Advances, J. Climate, 18, 1291–1314, 2005.

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.

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.

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.

Kutzbach, L., Thees, B., and Wilmking, M.: Identification of linear relationships from noisy data using errors-in-variables models - relevance for reconstruction of past climate from tree-ring and other proxy information, Climatic Change, 105, 155–177, 2011.

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.

Mu, Q., Jackson, C. S., and Stoffa, P. L.: A multivariate empirical-orthogonal-function-based measure of climate model performance, J. Geophys. Res., 109, D15101, doi:10.1029/2004JD004584, 2004.

Murphy, A. H.: Skill scores based on the mean square error and their relationships to the correlation coefficient, Mon. Weather Rev., 116, 2417–2424, 1988.

Murphy, A. H. and Epstein, E. S.: Skill scores and correlation coefficients in model verification, Mon. Weather Rev., 117, 572–581, 1989.

Murphy, M. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins M., and Stainforth, D. A.: Quantification of modelling uncertainties in a large ensemble of climate change simulations, Nature, 430, 768–772, 2004.

NRC – National Research Council: Surface temperature reconstructions for the last 2,000 years, The National Academies Press, Washington, D.C., 145 pp., 2006.

Osborne, C.: Statistical calibration: A review, Int. Stat. Rev., 59, 309–336, 1991.

Rowlands, D. J., Frame, D. J., Ackerley, D., Aina, T., Booth, B. B. B., Christensen, C., Collins, M., Faull, N., Forest, C. E., Grandey, B. S., Gryspeerdt, E., Highwood, E. J., Ingram, W. J., Knight, S., Lopez, A., Massey, N., McNamara, F., Meinshausen, N., Piani, C., Rosier, S. M., Sanderson, B. M., Smith, L. A., Stone, D. A., Thurston, M., Yamazaki, K., Yamazaki, Y. H., and Allen, M. R.: Broad range of 2050 warming from an observationally constrained large climate model ensemble, Nat. Geosci., 5, 256–260, 2012.

Stott, P. A., Allen, M. R., and Jones, G. S.: Estimating signal amplitudes in optimal fingerprinting, Part II: application to general circulation models, Clim. Dynam., 21, 493–500, 2003.

Tingley, M. P., Craigmile, P. F., Haran, M., Li, B., Mannshardt, E., and Rajaratnam, B.: Piecing together the past: statistical insights into paleoclimatic resonstructions, Quaternary Sci. Rev., 35, 1–22, 2012.

von Storch, H.: On the use of inflation in statistical downscaling, J. Climate, 12, 3505–3506, 1999.

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.