Special Issue Paper

Environmetrics

Received: 30 September 2011, Revised: 17 February 2012, Accepted: 27 February 2012, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/env.2142

Defining spatial comparison metrics for evaluation of paleoclimatic field reconstructions of the Common Era

Bo Li[a]* and Jason E. Smerdon[b]

Climate field reconstructions (CFR) of the Common Era (the last two millennia) provide important insights into the dynamics of past climate change that, in turn, have implications for the future. Multiple CFR methods have emerged in the literature, and comparisons between these methods using pseudoproxy experiments have been performed. These experiments, however, have not fully quantified the spatial skill of the CFRs, particularly with regard to the relative performance of each. Toward such ends, a formal statistical hypothesis test is proposed as a means of evaluating the differences between two random fields that integrate the differences in both the mean and the dependence structure. This involves a careful selection of the statistical model for the CFR residual process and systematic comparisons over different spatial scales. Application of this method yields a systematic assessment of the spatial character of five widely applied CFRs in a pseudoproxy experiment context. The analyses indicate that spatial differences among the five CFRs are not statistically significant. Further rigorous statistical assessments will help elucidate the strength and weakness of each CFR method, while quantifying the degree to which their spatial dissimilarities can be ascribed to methodological choices.

Keywords: climate field reconstruction; pseudoproxy experiment; random field comparison; paleoclimate

1. INTRODUCTION

Paleoclimatology offers a glimpse into historical climates over a range of timescales and mean states, while allowing us to test and inform our developing hypotheses about the climate system and climatic change (e.g., Schmidt, 2010). Among the many periods of relevance, the Common Era (CE) (the last two millennia) is an important target because high-resolution paleoclimatic proxies such as tree rings, corals, and ice cores are abundant enough over this interval to allow seasonal-to-annual reconstructions on regional to global spatial scales (e.g., Jones *et al.*, 2009). Studies of CE climate have already proved vital to understanding basic properties of the climate system and the vulnerabilities of present societies to climate variability that is not fully captured in modern observational records typically spanning the last 100–150 years. In just one such example, megadroughts of extreme duration and intensity not witnessed during the 20th century have been characterized throughout much of Mexico and the American Southwest (see Cook *et al.*, 2007 for a review), and similar drought vulnerabilities have been identified in areas of Asia and Indonesia (Cook *et al.*, 2010; Buckley *et al.*, 2010). Despite these and many other successes, however, there remain outstanding questions about our understanding of CE climate and its implications for the future.

Perhaps, one of the more widely debated areas of CE paleoclimatology involves the reconstruction of global or hemispheric temperatures by using networks of climate proxies derived from multiple proxy records (e.g., North *et al.*, 2006; Jansen *et al.*, 2007; Jones *et al.*, 2009; Stein, 2011). Although much progress has been made to understand the methods and data used to derive these reconstructions, there remain important unanswered questions about reconstruction uncertainties and their interpretations. These questions are tied to understanding the connections between climate and proxy responses across different spectral domains, the response of proxies to multiple environmental variables, the role of teleconnections and noise in calibration data, and the impact of specific proxy networks and methodological choices on derived reconstructions—questions that are ultimately fundamental to the success of efforts to reconstruct past climatic variability during the CE.

Among the different uncertainties that have been addressed in large-scale temperature reconstruction work, methodological assessments have become a recent focus (e.g., Jones *et al.*, 2009; Tingley *et al.*, 2012; Smerdon, 2012). An important distinction within these assessments is that reconstruction methodologies typically fall into two categories. The first involves index methods that target indices for reconstruction such as global or northern hemisphere (NH) mean temperatures, and the second comprises climate field reconstruction (CFR) methods that target hemispheric or global patterns, that is, spatial maps of temperature change. In general, reconstruction evaluations have focused on either index methods alone or on indices derived from large-scale composite averages of CFRs. Despite this focus on large-scale means, the

* *Correspondence to: Bo Li, Department of Statistics, Purdue University, West Lafayette, IN 47906, U.S.A. E-mail: boli@stat.purdue.edu*

a *Department of Statistics, Purdue University, West Lafayette, IN 47906, U.S.A*

b *Lamont-Doherty Earth Observatory of Columbia University, Palisades, NY, U.S.A*

spatial information offered by CFRs is perhaps the most valuable aspect of CE temperature reconstructions because of its ability to provide dynamical insights. For example, recent Asian drought CFRs have characterized how hydrological conditions in the region are affected by volcanic activity (Anchukaitis *et al.*, 2010), and a global CFR has helped characterize both hurricane variability in the Atlantic over the last millennium (Mann *et al.*, 2009b) and the response of the equatorial Pacific during the Medieval Climate Anomaly (Mann *et al.*, 2009a).

Despite the demonstrated utility of CFRs, it is not widely appreciated that very few hemispheric or global CFRs actually exist, although there are more regional CFRs for multiple climatic variables (e.g., Evans *et al.*, 2002; Cook and Krusic, 2004; Zhang *et al.*, 2004, Cook *et al.*, 2010, Neukom *et al.*, 2010). For example, only two of the twelve NH temperature reconstructions summarized in Assessment Report Four of the Intergovernmental Panel on Climate Change are derived from CFRs (Jansen *et al.*, 2007). Since the publication of the Assessment Report Four, only one additional global temperature CFR has been published (Mann *et al.*, 2009a). Research to produce large-scale temperature CFRs is thus still nascent. Their application can be complicated by the fact that they attempt to reconstruct more spatial locations than the number of temporal observations in the calibration interval (the regression problem is underdetermined) and are more likely dependent on the stability of climate-proxy connections, climate teleconnections, and proxy network distributions than index reconstructions. Much work remains to refine their methods and uncertainties, while expanding the proxy networks used to produce them. It therefore is critical to develop assessment metrics for evaluating the spatial information in derived CFR products and the degree to which one method may produce more spatial skill than another. The purpose of this manuscript is to explore and demonstrate the application of one such spatial assessment metric for use in evaluation of CFR products.

1.1. Pseudoproxy experiments

One important new tool for assessing the performance of CE reconstruction methodologies is millennium-length, forced transient simulations with fully coupled general circulation models (GCMs) (e.g., Gonzalez-Rouco *et al.*, 2003, 2006, 2011; Ammann *et al.*, 2007). These model simulations are used as test beds on which to evaluate the performance of a given reconstruction method by using controlled and systematic experiments—an approach that has come to be known as pseudoproxy experiments (PPEs) (see Smerdon (2011) for a review). The motivation for PPEs stems from the fact that real-world reconstructions are derived from many different methods, calibration choices, and proxy networks. Uncertainty in any given real-world reconstruction is therefore a combined result of the employed method, the adopted calibration data and calibration time interval, the spatial and temporal sampling of the proxy network, and the actual climate-proxy connection of each proxy record used for the reconstruction. If the objective is to isolate the impact of one of these factors, it is difficult to do so from comparisons between available real-world reconstructions. PPEs have allowed some of the aforementioned challenges to be circumvented by adopting a common framework that can be systematically altered and evaluated and thus test reconstruction methods and their dependencies.

The approach of PPEs is to extract a portion of a spatiotemporally complete GCM field in a way that mimics the available proxy and instrumental data used in real-world reconstructions. The principal experimental steps proceed as follows: (i) the complete GCM field is subsampled to approximate the availability of instrumental and proxy information; (ii) the time series that represent proxy information are added to noise series to simulate the temporal (and in some cases spatial) noise characteristics present in real-world proxy networks; and (iii) reconstruction algorithms are applied to the model-sampled pseudo "instrumental data" and pseudoproxy series to produce a reconstruction of the climate simulated by the GCM. The culminating fourth step is to compare the derived reconstruction with the known model target as a means of evaluating the skill of the applied method and the uncertainties expected to accompany a real-world reconstruction product.

Evaluations of CFRs have been performed in multiple PPEs, but very few of them have focused specifically on the spatial skill of the reconstructions. Some studies have reported summaries of field statistics or provided spatial plots of limited assessment metrics (Rutherford *et al.*, 2003; Mann *et al.*, 2005, 2007), but the primary evaluations of CFR methods in PPEs to date have focused on their ability to derive skillful NH or global mean indices. Such evaluations are insufficient for assessing the spatial performance of CFRs (Smerdon *et al.*, 2011a). The few PPEs that have directly assessed spatial skill of CFR methods have reported significant variations in regional performance (e.g., Riedwyl *et al.*, 2009; Tingley and Huybers, 2010a,b; Smerdon *et al.*, 2008a, 2011a,b). Current assessments of CFR spatial performance indicate the need to more fully vet the field skill of contemporary methods by using multiple models and scenarios, while more directly connecting PPE results to the specific characteristics of real proxies and climate fields. Rigorous comparisons based on spatial performance will help further elucidate the strength and weakness of each CFR method, while quantifying the degree to which spatial dissimilarities in CFRs can be ascribed to methodological choices.

1.2. Statistical developments in comparing random processes

Pseudoproxy experiments are the context in which we consider the spatial performance of a suite of CFRs derived from the collection of multivariate linear regression methods typically employed for reconstruction of CE climate. A principal question that we address is whether different CFRs can be considered statistically distinct and therefore justify the different methodological choices that have been used to derive them. By definition, statistical methods are required to address this question because two realizations governed by the same underlying space–time model can appear very different because of the randomness of noise. For example, the well-known two-sample t test is used to evaluate whether two different samples have essentially the same mean after randomness is taken into account. Many examples have shown that spurious conclusions can be drawn if noise is not appropriately considered (e.g., Carroll *et al.*, 2006).

The mean squared error (MSE), which is the mean squared difference of two random fields over all locations, is a typical statistical measure used to quantify the similarity between two sets of observations over all locations. Nevertheless, this measure can potentially be dominated by noise and unable to recover the true relationship between two fields. A large value of MSE would only indicate that two random fields are *obviously* different but leaves unknown whether dissimilarities are caused by noise or underlying field structures. This is analogous to the

argument that two samples with distinct sample averages may indeed share the same population mean. Other types of univariate summaries such as the root MSE and the anomaly correlation coefficient (e.g., Briggs and Levine, 1997) are also typically not sufficient for descriptions of the differences between fields. Given this discussion, it might appear that the Hotelling T^2 test (see Mardia *et al.*, 1979) could be used to compare two spatial reconstructions in this context, but it is important to note that this test compares two multivariate samples rather than two stochastic processes. Because either the reconstructions at spatial grids or the observations at irregularly spaced monitoring sites are just realizations of an underlying continuous process of climate, it is more appropriate to compare their inherent structures rather than a few observations embedded in the process. Moreover, it is unrealistic to treat the reconstructions at different times as replicates. Therefore, even if the spatial observations are treated as multivariate data, the Hotelling T^2 is not applicable because the approach relies on replication to derive estimates of the mean and covariance estimator.

Discussions in the literature that are closely related to this topic are focused mainly on comparisons of the prediction accuracy of different forecasts. Briggs and Levine (1997), Shen *et al.* (2002), and Pavlicová *et al.* (2008) compared two fields on the basis of the wavelet transform but only for gridded random fields for which the grid sizes must be a dyadic power. Diebold and Mariano (1995) evaluated whether two time series forecasts are equally accurate by examining whether their loss differential is essentially zero on average, where the loss function can be MSE or any other linear or nonlinear function of prediction errors. Snell *et al.* (2000) and Wang *et al.* (2007) later applied this idea in a spatial setting. Other forms of point estimates of prediction accuracy or loss functions have also been used (e.g., Atger, 2003; Gong *et al.*, 2003; Willis, 2002). All these previous methods ignored either the uncertainty associated with the estimates of loss functions or the potential spatial dependence in those estimates. Very recently, Hering and Genton (2011) developed a hypothesis test to evaluate the difference between two random fields in terms of user-specified loss functions. Their test for the first time considered the spatial correlation of the loss differential in estimating its uncertainty, but it still did not consider the difference in dependence structures.

A spatial or spatio-temporal random field is characterized by both the first-order and higher-order moments. For example, a Gaussian random field is determined by the first two moments. Missing one component in the test fails to provide a thorough comparison between two random fields. Lund and Li (2009) proposed a distance metric between two time series that integrate the differences in both the mean and covariance structures. This cannot be directly applied to the comparison between two random fields because the time series has different characteristics and properties than random fields due to its natural order in time. We propose a testing approach that combines the discrepancies in both the mean and covariance structure to assess the difference between two random fields. For a Gaussian random process, the null hypothesis indicates that the random fields are simply two realizations from the same underlying structure, and their differences are only due to noise. Discrepancies either in mean or in covariance structure will lead to the rejection of the test. The hypothesis that we attempt to test is different from Shen *et al.* (2002) and Pavlicová *et al.* (2008) who investigated the equality of only the first moment although their investigation has the second moment taken into consideration. Our aim is also different from Krishnaiah *et al.* (1980) who addressed the simultaneous hypothesis test for either the mean vector or the covariance matrix for multiple random processes. Their focus is on how to control the family error rate for the multiple testing, whereas the main challenge in our work is to combine both the mean and covariance matrix in a single hypothesis testing.

In Section 2, we describe the provenance of the PPE data used in our analysis. In Section 3, we describe the new hypothesis testing approach and a simulation study for evaluating this method. In Section 4, we extensively analyze and compare CFRs with five reconstruction methods by using the new testing approach. We end with Section 5, which provides a discussion of the testing method and the phenomena we observed in CFR comparisons. All the data are publicly accessible, and all the codes related to this article are available upon request.

2. DATA

We use a pseudoproxy framework derived from the millennial simulation (850–1999 CE) of the National Center for Atmospheric Research Community Climate System Model, version 1.4, a coupled atmosphere–ocean GCM that has been driven with natural and anthropogenic forcings (Ammann *et al.*, 2007). The annual means of the modeled temperature field have been interpolated from the native resolution to an even 5° latitude–longitude grid and comprise the grid from which all samplings are performed (Smerdon *et al.*, 2008b). The interpolated model field that we employ does not suffer from processing problems previously described by Smerdon *et al.* (2010).

Pseudoproxies are sampled from the 104 temperature grid points that approximate the actual proxy locations in the most populated nest of the proxy network by Mann *et al.* (1998; hereinafter Mann–Bradley–Hughes (MBH98); all pseudoproxies are taken as available for the entire reconstruction interval. See Figure 1 for the locations of the temperature time series used as pseudoproxies to derive all of the CFRs used herein. Each of the sampled pseudoproxy series is perturbed at two Gaussian white noise levels: signal-to-noise ratios (SNRs) of 0.5 and 0.25, by standard deviation. Most PPE studies use no-noise pseudoproxy networks as baseline predictor networks (see Smerdon (2011) for a review). Various colors of noise with different levels of variance are then added to the sampled temperature time series to establish collections of pseudoproxy networks with multiple SNRs. The most widely applied choice of noise has been Gaussian white noise at SNR values of 1.0, 0.5, and 0.25 by standard deviation. It is generally argued that SNRs on the order of 0.5–0.25 (by standard deviation) are representative of the actual noise level in real-world proxy records, which motivates our choice earlier. This estimate for real-world noise levels is based on correlations between collocated instrumental temperatures and proxies during their interval of overlap. Despite matching these estimates, real-world proxy records have signal and noise characteristics that are unique to the physical, chemical, or biological system from which they are measured. None of these characteristics is fully captured in currently employed PPEs, making their representations of signal and noise perhaps the largest idealization.

The targeted mean annual temperature field in the Community Climate System Model data has been spatially masked to approximate the availability of instrumental temperature data (Mann *et al.*, 2008; Smerdon *et al.*, 2011) by excluding grid points missing more than 30% of the annual data between 1856 and 1998 CE in the dataset by Brohan *et al.* (2006). The resulting field has a total of 1732 grid cells (Mann *et al.*, 2008); missing grid cells are indicated in subsequent figures as white grid cells. All tested methods are calibrated from 1856 to 1980

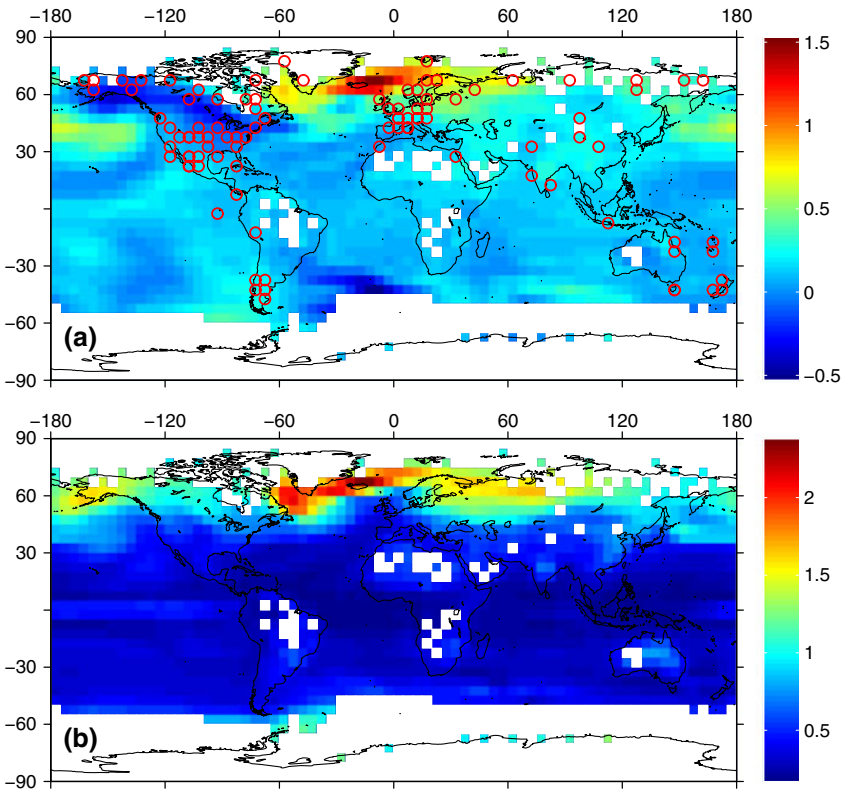

Figure 1. (a) Temporal mean and (b) temporal standard errors of residual process from canonical correlation analysis. Grid-box locations of the temperature time series used in the pseudoproxy climate field reconstruction experiments are shown in the upper panel. Locations approximate the distribution of the Mann–Bradley–Hughes (MBH98) network

CE, and all validation statistics are calculated during the reconstruction interval from 850 to 1855 CE. The adopted experimental setup can be considered a best-case scenario for real-world conditions, whereas additional modifications to the PPE framework to more fully mimic real-world proxies will only degrade the results (e.g., von Storch *et al.*, 2004, 2006; Mann *et al.*, 2007). This construction nevertheless has been used by a wide range of pseudoproxy studies (Smerdon, 2012).

By using the aforementioned framework, temperature fields derived from five different CFR methods are used in our analysis: inverse ordinary least squares regression (MBH hereinafter), regularized expectation maximization using truncated total least squares (TTL hereinafter), TTL using a hybrid spectral calibration scheme split at the 20-year period (TTH hereinafter), ridge regression (RDG hereinafter), and canonical correlation analysis (CCA hereinafter). The MBH method was applied by Mann *et al.* (1998) as emulated by von Storch *et al.* (2006). TTL and TTH CFRs were performed as described and advocated by Mann *et al.* (2007) and Rutherford *et al.* (2010). Standard RDG(Hoerl and Kennard, 1970) was used for the RDG CFRs, using minimization of the generalized cross validation function for ridge parameter selection (Golub *et al.*, 1979). The CCA method was applied as described in Smerdon *et al.* (2011b). Further description and characterization of the reconstructed fields and associated methods can be found in Smerdon *et al.* (2011a), and the data can be publicly accessed at http://www.ldeo.columbia.edu/~jsmerdon/2011_grl_supplement.html.

3. A NEW METHOD TO ASSESS THE DISCREPANCIES BETWEEN RANDOM FIELDS

3.1. Hypothesis testing

We propose a formal test that can detect whether two random fields are characterized by the same first and second moments. This test is in the same spirit of Lund and Li (2009) who investigated a distance measure that integrates the differences in the underlying mean and covariance structure between two time series, whereas our test is developed in a spatial or spatio-temporal context for which the time series framework cannot be applied. Let $\mathbf{x}$ be the spatial (spatio-temporal) locations for the observations. We have $\mathbf{x} = \mathbf{s}$ in a spatial context and $\mathbf{x} = (\mathbf{s}, t)$ for a spatio-temporal random field, where $\mathbf{s}$ denotes a spatial location and t is time. Let $Y(\mathbf{x})$ denote a random field observed over $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in a spatial domain D. A standard modeling framework for $Y(\mathbf{x})$ is

$$Y(\mathbf{x}) = \mu(\mathbf{x}) + \epsilon(\mathbf{x})$$

where $\mu(\mathbf{x})$ is the spatially varying mean of $Y(\mathbf{x})$ and $\epsilon(\mathbf{x})$ is a spatially correlated error for which $(\epsilon(\mathbf{x}_1), \ldots, \epsilon(\mathbf{x}_n))^{\mathrm{T}} \sim F(\mathbf{0}, \mathbf{\Omega})$ for an n-variate distribution function F and a covariance matrix $\mathbf{\Omega}$. For instance, F can be multivariate normal or multivariate t_ν with degrees

of freedom $\nu > 2$. The trend $\mu(\mathbf{x})$ can either take a linear or nonlinear function of covariates or a nonparametric smoothing function such as a spline, and the error $\epsilon(\mathbf{x})$ is allowed to be nonstationary. Let $C(\mathbf{x}, \mathbf{x}+\mathbf{h}) = \text{cov}\{Y(\mathbf{x}), Y(\mathbf{x}+\mathbf{h})\}$, $\mathbf{y} = (Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n))^{\mathrm{T}}$, $\boldsymbol{\mu} = (\mu(\mathbf{x}_1), \ldots, \mu(\mathbf{x}_n))^{\mathrm{T}}$, and $\boldsymbol{\epsilon} = (\epsilon(\mathbf{x}_1), \ldots, \epsilon(\mathbf{x}_n))^{\mathrm{T}}$ and thus $\boldsymbol{\Omega} = \text{var}(\boldsymbol{\epsilon})$. Now suppose we have another random process, $Y'(\mathbf{x})$, for which we can correspondingly define $C'(\mathbf{x}, \mathbf{x}+\mathbf{h})$, $\mathbf{y}'$, $\boldsymbol{\mu}'$, $\boldsymbol{\epsilon}'$, and $\boldsymbol{\Omega}'$ by using $Y'(\mathbf{x})$.

Our goal is to assess whether $Y(\mathbf{x})$ and $Y'(\mathbf{x})$ share a common mean and correlation structure, or equivalently, whether their discrepancies are mainly due to the randomness of the particular observations. This question can be formulated into a hypothesis test, where the null hypothesis is

$$H_0 : \mu(\mathbf{x}) = \mu'(\mathbf{x}) \text{ and } C(\mathbf{x}, \mathbf{x}+\mathbf{h}) = C'(\mathbf{x}, \mathbf{x}+\mathbf{h}) \tag{3.1}$$

for any $\mathbf{x} \in D$ and any $\mathbf{h}$ such that $\mathbf{x}+\mathbf{h} \in D$. Assume we observe $Y(\mathbf{x})$ and $Y'(\mathbf{x})$ at the same set of locations, as for two given CFRs. Under the null hypothesis, $\boldsymbol{\mu} = \boldsymbol{\mu}'$ and $\boldsymbol{\Omega} = \boldsymbol{\Omega}'$, and $\tilde{\mathbf{y}} = \boldsymbol{\Omega}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$ and $\tilde{\mathbf{y}}' = \boldsymbol{\Omega}^{-1/2}(\mathbf{y}' - \boldsymbol{\mu})$ yield two uncorrelated samples. These two uncorrelated samples both follow $N(0,1)$ if F is n-variate normal, and both follow the central t_ν distribution if F is n-variate t_ν with degrees of freedom $\nu > 2$ (see Kotz and Nadarajah, 2004). Note $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ can be replaced by $\boldsymbol{\mu}'$ and $\boldsymbol{\Omega}'$ in computing $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{y}}'$. A violation of either component in H_0 will result in unequal distributional properties of $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{y}}'$. Thus, for these two common multivariate distributions, we can evaluate the components in the null hypothesis by examining the equality of the univariate distribution of $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{y}}'$. This can be accomplished by the two-sample Kolmogorov–Smirnov (KS) test. The two-sample KS test quantifies the distance between the empirical distribution functions of two samples, and the null distribution of the KS test statistic is calculated under the null hypothesis that the samples are drawn from the same distribution. This testing approach integrates the two components in the null hypothesis into one single test.

Both $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ in (3.1) are typically unknown and need to be estimated. In general, if there is a priori information about the cause or the approximate pattern of the trend, we will choose a parametric form for $\mu(\mathbf{x})$. Otherwise, a nonparametric form may be preferred. In the latter case, there is a risk of either inadequate estimation or overfitting of $\mu(\mathbf{x})$ without the knowledge of $\boldsymbol{\Omega}$ (e.g., Opsomer *et al.*, 2001; Francisco-Fernandez and Opsomer, 2005) because the variation in a random process can either be caused by the trend or by the covariance. The challenge of estimating $\mu(\mathbf{x})$ and associated solutions are discussed in detail in Hering and Genton (2011) and Bliznyuk *et al.* (2012). Nevertheless, if a given data set has specific features that can provide additional information about the mean and covariance, such as the temporal stationarity of CFR residual processes discussed in Section 4, they can be exploited as a means of estimating $\mu(\mathbf{x})$. The estimation of $\boldsymbol{\Omega}$ can be carried out by fitting a parametric covariance model such as exponential, Matérn, and so on. An alternative procedure is to use a nonparametric method to estimate the covariance structure (e.g., Huang et al., 2008). Once $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Omega}}$ are estimated, the estimation of $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{y}}'$ by $\widehat{\tilde{\mathbf{y}}} = \widehat{\boldsymbol{\Omega}}^{-1/2}(\mathbf{y} - \widehat{\boldsymbol{\mu}})$ and $\widehat{\tilde{\mathbf{y}}'} = \widehat{\boldsymbol{\Omega}}^{-1/2}(\mathbf{y}' - \widehat{\boldsymbol{\mu}})$ naturally follows. The KS test is subsequently applied to $\widehat{\tilde{\mathbf{y}}}$ and $\widehat{\tilde{\mathbf{y}}'}$ to detect whether they have equal empirical distribution functions.

Note that if Y and Y' are realizations at two sets of different locations, say Y represents the real temperatures at monitoring sites and Y' the CFR at spatial grids, the locations of Y do not necessarily overlap with those of Y'. In this case, the hypothesis test is still applied because the null hypothesis in (3.1) simply implies that both $\mu(\mathbf{x})$ and $\mu'(\mathbf{x})$ follow an identical parametric or nonparametric form, and $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}'$ are equally governed by the same correlation structure, but to obtain $\widehat{\tilde{\mathbf{y}}'}$, the estimated parametric or nonparametric form of $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ must be used to compute $\widehat{\boldsymbol{\mu}}'$ and $\widehat{\boldsymbol{\Omega}}'$ for $\mathbf{y}'$, and then $\widehat{\tilde{\mathbf{y}}'} = \widehat{\boldsymbol{\Omega}'}^{-1/2}(\mathbf{y}' - \widehat{\boldsymbol{\mu}}')$.

3.2. Simulation studies

Simulations are conducted to evaluate the sizes and powers of the proposed test, particularly for various types of alternative hypotheses that may violate either component in the null hypothesis. Because the mean and inhomogeneous variance in our data set can be estimated easily by taking advantage of the temporal stationarity of the data, our simulation experiment will not involve varying mean function and inhomogeneous variances but rather evaluate the main idea of the test. We generate $\mathbf{y}_1 \sim N(\mu_1 \mathbf{1}_n, \boldsymbol{\Omega}_1)$ and $\mathbf{y}_2 \sim N(\mu_2 \mathbf{1}_n, \boldsymbol{\Omega}_2)$, where $\mathbf{1}_n$ is a vector of one with length n, $\boldsymbol{\Omega}_1$ is governed by the covariance function $C_1(\mathbf{h}) = \sigma_1^2 \exp(-||\mathbf{h}||/\phi_1)$, and $\boldsymbol{\Omega}_2$ by $C_2(\mathbf{h}) = \sigma_2^2 \exp(-||\mathbf{h}||/\phi_2)$. Both are exponential covariance models but with possibly different variance parameters σ_1^2 and σ_2^2 and different range parameters ϕ_1 and ϕ_2. The value of the parameters and the sample sizes are listed in Table 1. The null hypothesis for this setting can be written as H_0: $\mu_1 = \mu_2$, $\sigma_1^2 = \sigma_2^2$, and $\phi_1 = \phi_2$, which defines the null subspace of $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \phi_1, \phi_2)^{\mathrm{T}}$ in $\mathbb{R}^6$. The alternative space contains an array of subspaces in $\mathbb{R}^6$ determined by the permutation of changing "=" to"$\neq$" for each of the three components in H_0.

We compute the probability of rejecting H_0 at various subspaces of parameters over 1000 simulation runs. In testing the null hypothesis, we consider only a sequence of different decay rates of the correlation as the processes are usually centered and scaled before performing the test. We examine the powers for each of the unequal mean, unequal variance and unequal range parameters separately. We also evaluate the sizes and powers for different sample sizes. The maximum likelihood method is employed to estimate the parameters in the model. Table 1 shows that the sizes are close to the nominal level 0.05. The powers increase rapidly as either μ_1 and μ_2 or σ_1^2 and σ_2^2 diverge yet less rapidly when ϕ_2 departs from ϕ_1. The powers also become larger as the sample size increases, as expected. Table 1 also reports the powers for deviations from the null hypothesis at different degrees when all three components in H_0 take the "$\neq$" sign, and it can be seen that the powers are lifted as all three parameters involved in $\mathbf{y}_2$ are further apart from those involved in $\mathbf{y}_1$.

We additionally perform two extra simulation experiments to evaluate the effect of covariance estimation on sizes and powers. One employs the same simulation settings as described earlier, except that it replaces the estimated covariance function by its true covariance function in computing the test statistic. The comparison between the results from this simulation and Table 1 indicates that the uncertainty in the covariance estimation has negligible impacts on the sizes and powers. The second additional experiment also follows the basic settings in the original simulation, but here the spatial processes are generated using a Matérn model, $C(\mathbf{h}) = \sigma^2(\alpha\|\mathbf{h}\|)^\nu K_\nu(\alpha\|\mathbf{h}\|)/\{\Gamma(\nu)2^{\nu-1}\}$, with

Table 1. Empirical sizes and powers of testing the similarity between two random fields

		Parameters						Sample size		
	Parameter subspace	μ_1	μ_2	σ_1^2	σ_2^2	ϕ_1	ϕ_2	10×10	15×15	20×20
Size	$\mu_1 = \mu_2$	0	0	1	1	2	2	0.045	0.063	0.051
	$\sigma_1^2 = \sigma_2^2$	0	0	1	1	3	3	0.048	0.064	0.047
	$\phi_1 = \phi_2$	0	0	1	1	4	4	0.055	0.070	0.050
		0	0	1	1	5	5	0.052	0.067	0.056
Power	$\mu_1 \neq \mu_2$	0	1	1	1	2	2	0.393	0.618	0.777
	$\sigma_1^2 = \sigma_2^2$	0	2	1	1	2	2	0.856	0.977	0.998
	$\phi_1 = \phi_2$	0	3	1	1	2	2	0.872	1.000	1.000
		0	4	1	1	2	2	0.996	1.000	1.000
	$\mu_1 = \mu_2$	0	0	1	2	2	2	0.197	0.541	0.863
	$\sigma_1^2 \neq \sigma_2^2$	0	0	1	3	2	2	0.534	0.973	1.000
	$\phi_1 = \phi_2$	0	0	1	4	2	2	0.795	1.000	1.000
		0	0	1	5	2	2	0.933	1.000	1.000
	$\mu_1 = \mu_2$	0	0	1	1	2	3	0.111	0.200	0.268
	$\sigma_1^2 = \sigma_2^2$	0	0	1	1	2	4	0.223	0.468	0.727
	$\phi_1 \neq \phi_2$	0	0	1	1	2	5	0.380	0.761	0.961
		0	0	1	1	2	6	0.504	0.923	0.996
	$\mu_1 \neq \mu_2$	0	1	1	2	2	3	0.422	0.643	0.820
	$\sigma_1^2 \neq \sigma_2^2$	0	2	1	3	2	4	0.775	0.950	0.995
	$\phi_1 \neq \phi_2$	0	3	1	4	2	5	0.909	0.991	0.999
		0	4	1	5	2	6	0.962	0.999	1.000

Nominal level is 0.05.

the smoothness parameter ν being 1.0, 1.5, and 2.0, respectively. In this model, $\Gamma(\cdot)$ is the gamma function, and $K_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν. The parameter α represents the spatial scale controlling the decaying rate of the covariance function. Despite the fact that the Matérn model is used in the data generation, we still fit an exponential covariance function to the data. Table 2 shows the sizes and powers of the test corresponding to each of the three smoothness parameters. The results imply that a misspecified model can deteriorate all the sizes and the powers associated with different means. The sizes become worse when the smoothness parameter of the Matérn model is further apart from 0.5, which corresponds to the exponential covariance function. These observations indicate that we should be cautious about the model selection to retain the accurate size and high power.

4. ASSESSMENT OF DISCREPANCIES AMONG DIFFERENT CLIMATE FIELD RECONSTRUCTIONS

We use the previously described and evaluated hypothesis test to characterize differences between the five CFRs introduced in Section 2. Our driving question is the degree to which the CFRs can be considered spatially distinct. We perform all of the tests on the basis of the residual process obtained by taking the difference between each CFR and the target. Because the target is "error free," while the CFR carries noise because of the noise-added pseudoproxies, the comparison between residual processes is equivalent to a comparison between CFRs.

4.1. A statistical model for the residual process

We use the CCA CFR based on SNR = 0.5 pseudoproxies to illustrate the specific features of the residual process, but the CFRs from the other four methods display similar patterns, similarly so for experiments using the five CFRs derived from pseudoproxies with a reduced SNR of 0.25. To illustrate the spatial dependence of the CCA residual process, we present its temporal mean and temporal standard errors in Figure 1 where the residuals are seen to be maximized in the North Atlantic region and vary more widely in the high northern latitude areas. The variations with respect to latitude are associated with the known increased variability associated with the poles relative to the tropics and thus give rise to the larger residuals in the polar regions. Secondary structures are, of course, dependent on additional factors, such as the spatial sampling of the pseudoproxies and teleconnections in the target field. Figure 1 strongly supports the presence of a nonconstant mean and heterogeneous variance in the residual process.

Table 2. Empirical sizes and powers of testing the similarity between two random fields with misspecified covariance model. The true model is Matérn covariance function with smoothness parameter ν, whereas the fitted model is exponential covariance function

		Parameters						Smoothness parameter		
	Parameter subspace	μ_1	μ_2	σ_1^2	σ_2^2	ϕ_1	ϕ_2	1	1.5	2
Size	$\mu_1 = \mu_2$	0	0	1	1	2	2	0.007	0.004	0.045
	$\sigma_1^2 = \sigma_2^2$	0	0	1	1	3	3	0.007	0.049	0.244
	$\phi_1 = \phi_2$	0	0	1	1	4	4	0.017	0.150	0.469
		0	0	1	1	5	5	0.035	0.314	0.647
Power	$\mu_1 \neq \mu_2$	0	1	1	1	2	2	0.109	0.055	0.112
	$\sigma_1^2 = \sigma_2^2$	0	2	1	1	2	2	0.464	0.227	0.304
	$\phi_1 = \phi_2$	0	3	1	1	2	2	0.805	0.523	0.621
		0	4	1	1	2	2	0.942	0.777	0.872
	$\mu_1 = \mu_2$	0	0	1	2	2	2	0.415	0.412	0.474
	$\sigma_1^2 \neq \sigma_2^2$	0	0	1	3	2	2	0.958	0.927	0.904
	$\phi_1 = \phi_2$	0	0	1	4	2	2	1.000	0.996	0.991
		0	0	1	5	2	2	1.000	1.000	0.999
	$\mu_1 = \mu_2$	0	0	1	1	2	3	0.379	0.720	0.836
	$\sigma_1^2 = \sigma_2^2$	0	0	1	1	2	4	0.974	0.998	0.997
	$\phi_1 \neq \phi_2$	0	0	1	1	2	5	1.000	1.000	1.000
		0	0	1	1	2	6	1.000	1.000	1.000
	$\mu_1 \neq \mu_2$	0	1	1	2	2	3	0.148	0.137	0.295
	$\sigma_1^2 \neq \sigma_2^2$	0	2	1	3	2	4	0.474	0.535	0.768
	$\phi_1 \neq \phi_2$	0	3	1	4	2	5	0.720	0.850	0.935
		0	4	1	5	2	6	0.863	0.959	0.982

Nominal level is 0.05.

A particular spatial model with varying mean and nonstationary errors tailored for the residual process at each time point is hence proposed as

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + \tau(\mathbf{s})Z(\mathbf{s}) \tag{4.1}$$

where $Z(\mathbf{s})$ is sampled from $N(\mathbf{0}, \boldsymbol{\Sigma})$ for a covariance matrix $\boldsymbol{\Sigma}$ and the Gaussian approximation of the error distribution is justified by examining the normality of the scaled residuals after model fitting. Model (4.1) allows the mean $\mu(\mathbf{s})$ and variance $\tau(\mathbf{s})$ to be heterogeneous in the spatial domain as indicated by the characteristics of the residual process.

We estimate $\mu(\mathbf{s})$ and $\tau(\mathbf{s})$ as the sample mean and standard deviation from each location-specific time series, and then we consider $Z(\mathbf{s})$ as an approximately stationary spatial processes with zero mean and homogeneous variance. We present two scaled residual processes of CCA at year 1150 and 1450 in Figure 2 for illustration. These two plots show that the scaled processes are free of the pattern as displayed in Figure 1 (a), but they are spatially correlated and thus require a spatial covariance model to account for their dependency structures. The temporal correlation is not visually as strong as the spatial correlation. To formally assess the temporal correlation, we apply the Ljung–Box test on the time series $Z(\mathbf{s})$ at each spatial grid. We find that the time series at most of the grid points exhibit certain correlations. Consequently, we examined their correlation structures by the autocorrelation function and partial autocorrelation function. Two together imply that an autoregressive model of order one would be sufficient. We therefore have tried prewhitening each time series before performing the test. We nevertheless find that the conclusion of our testing results remains the same.

4.2. Testing results

A regionally resolved approach allows us to locate where the possible significant differences occur on the basis of the applied statistical hypothesis test and also to avoid the computational challenge associated with large spatial data. We therefore divide the global reconstruction domain into 10 regions as shown in Figure 3 to apply our assessment regionally. Each region contains roughly 170 spatial grids to retain power while still maintaining the flexibility to examine the local differences between CFRs. For each region, we assume $Z(\mathbf{s})$ to be a second-order stationary stochastic process with an exponential covariance model, $\text{cov}\{Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h})\} = \sigma^2 \exp(-\|\mathbf{h}\|/\phi)$, where σ^2 and ϕ are called variance and range parameters, respectively. We also investigated the use of the Matérn rather than the exponential covariance function in our error models to account for the smoothness of the stochastic process. North *et al.* (2011) suggested a Matérn model of order one for the

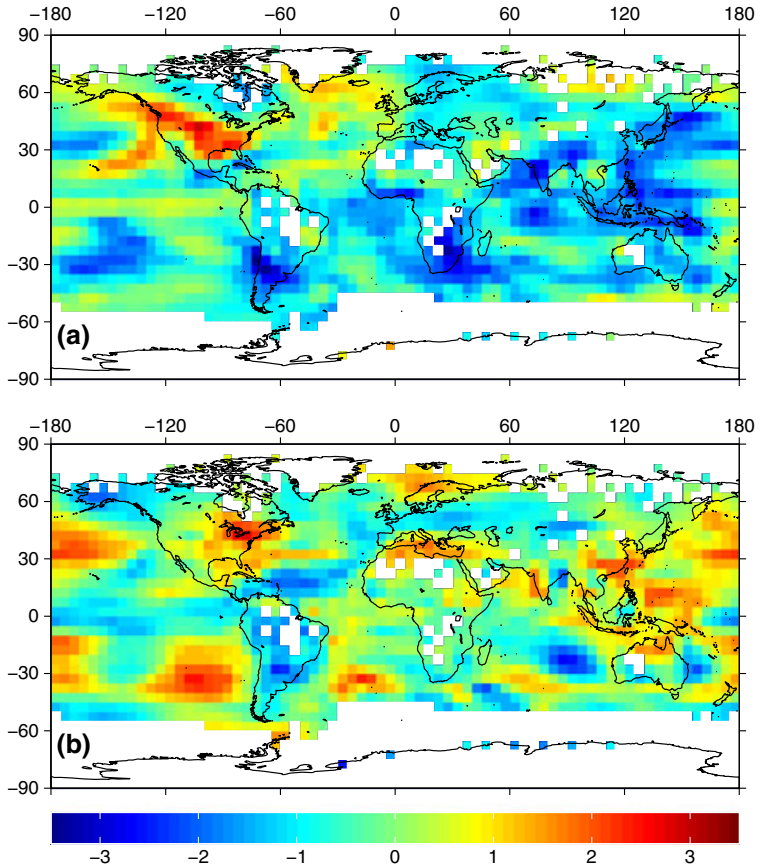

Figure 2. Scaled canonical correlation analysis residual processes at year (a) 1150 and (b) 1450

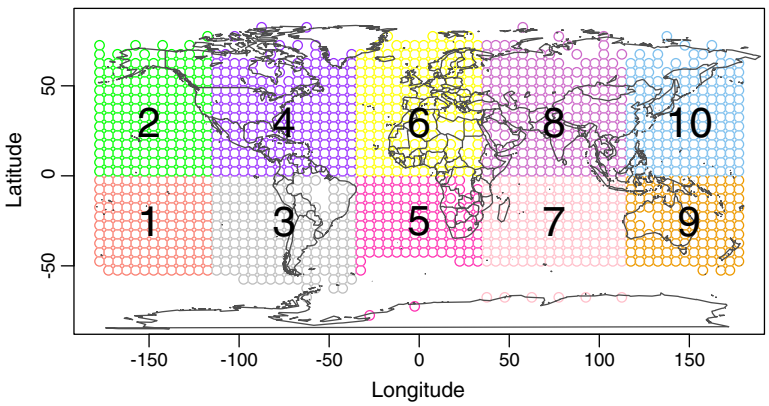

Figure 3. Regional divisions of the global spatial domain for the 10-region case considered in the main text

temperature process, but there is no evidence in our data to favor this choice. Given the difficulty of estimating smoothness parameters in the Matérn class (Stein, 1999; p. 219), we simply tried Matérn covariance functions of either order one or order two in the analysis. We found no appreciable differences using this more flexible covariance model. For the convenience of model fitting, we therefore keep the exponential covariance function in our error model and use the maximum likelihood method based on geographic distance between locations to estimate their parameters.

By using the previously defined conventions for the 10-region configuration, we obtain the p-values for each region over time. Figure 4 reports the overall average p-values for each pairwise comparison of the five derived CFRs. The comparison between CCA and MBH has the highest average p-values, whereas the comparison between RDG and TTH has the lowest. In general, TTH corresponds to the lowest p-values and TTL to the second lowest among all the pairs associated with any of the CCA, MBH, and RDG CFRs, although p-values between TTH and TTL are relatively high. This latter result is expected given that TTH and TTL apply the same method and only differ in that TTH applies the hybrid spectral calibration approach. Moreover, TTH is mostly distinct from all other CFRs (with the exception of

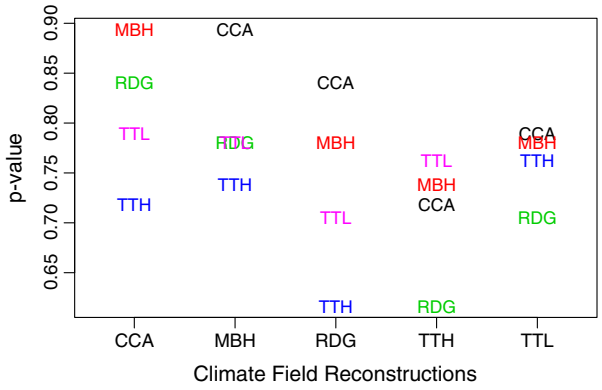

Figure 4. Overall average *p*-values of the pairwise comparisons between five climate field reconstructions

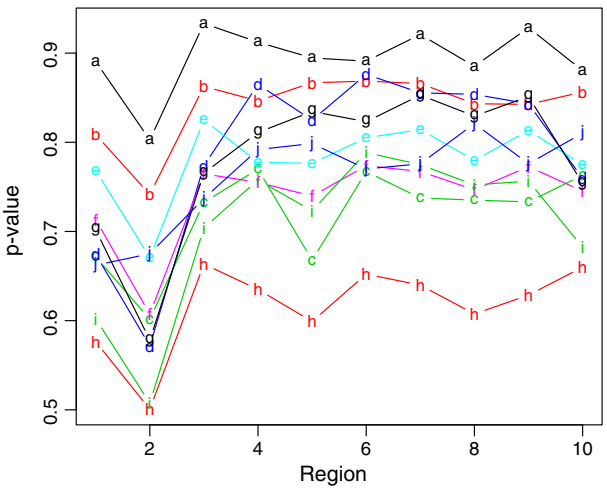

Figure 5. Regional average *p*-values for 10 regions. The 10 pairwise comparisons are indexed by the following: a, CCA–MBH; b, CCA–RDG; c, CCA–TTH; d, CCA–TTL; e, MBH–RDG; f, MBH–TTH; g, MBH–TTL; h, RDG–TTH; i, RDG–TTL; j, TTL–TTH

TTL), indicating that the hybrid calibration may be the most discriminating methodological choice among the applied methods. In contrast, CCA tends to have the largest *p*-values when it is compared with the MBH, RDG, and TTL CFRs, thus suggesting that it shares the most common features across all of the derived CFRs. All of the average *p*-values are nevertheless very large, being above 0.6 and in most cases larger than 0.7. These average assessments thus indicate that any differences among the derived CFRs are not statistically significant.

Despite the global assessment presented in Figure 4, it is still possible that regional differences among the CFRs may exist. We therefore investigate the *p*-values associated with each individual region in Figure 5, which plots the average regional *p*-values of all pairwise comparisons in each region. Although some of the regions yield *p*-values somewhat smaller than the global averages, they again are large and only reach slightly below 0.5 for some CFR comparisons in region 2. Regional variability across methods is indicated by variable *p*-value ranking in each region, but in general, the 10 comparisons are bounded by the CCA–MBH and RDG–TTH pairs yielding maximum and minimum *p*-values, respectively. This mirrors the global results shown in Figure 4.

Nine of the 10 pairwise comparisons consistently have region 2 corresponding to their lowest *p*-values among all the regions, indicating that CFRs in this region may exhibit more differences than the other areas. This variability is fully described in Figure 6, in which *p*-value boxplots are shown specifically for region 2 for each of the 10 pairwise comparisons. This figure shows that the RDG–TTH and RDG–TTL pairs have relatively low *p*-values, whereas the CCA–MBH pair again indicates larger *p*-values than other combinations. Some individual years can also reach smaller *p*-values, as indicated by the lower bound of the boxplots. Although there are a few statistically significant values at certain years, these analyses once again indicate that the differences among the five CFRs that we consider are not broadly significant even when individual regions and years are considered.

To test the possible effect of regional size on our testing results, we also have partitioned the entire spatial domain into only three regions by evenly splitting the globe by longitude. All of the aforementioned analyses were repeated for this three-region configuration. These new regions contained about 600 spatial grids and were expected to have higher power for the test than a region that only contains 170 spatial grids, that is, the 10-region configuration used earlier. Nevertheless, we found that results were very similar to those obtained from the 10 regions. In further sensitivity investigations, we have also performed the analysis for CFRs derived from SNR $= 0.25$ pseudoproxies because all the aforementioned results are based on pseudoproxies with SNRs of 0.5. These tests confirmed that the pattern of *p*-values for the CFRs with SNR $= 0.25$ is consistent with those we have shown for the SNR $= 0.5$ case.

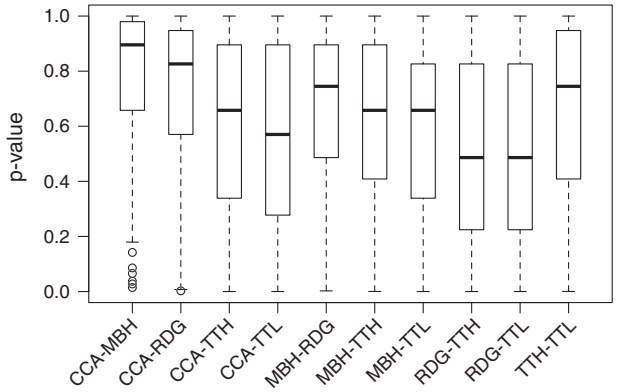

Figure 6. *p*-Values of pairwise comparisons over region 2

5. DISCUSSION AND CONCLUSIONS

Motivated by the scientific question of evaluating different CFR methods, this investigation has accomplished two complementary goals: (i) it has proposed a new method for assessing the spatial dissimilarities between random fields; and (ii) it has used the method to assess the statistical differences between CFRs derived from five different multivariate linear regression techniques in the context of PPEs. The proposed test focuses on the differences in both the mean and covariance structure between two random fields and integrates those two components into one single test for the assessment of their discrepancies. It is easy to implement, and the simulations verified an accurate size and expected power of the test. The new statistical test is expected to have further application in other disciplines. Examples include studies of regime transitions in solid-earth geophysics, atmospheric and oceanographic dynamics, and detection of dynamical changes in complex systems in physical, medical, engineering, and economic sciences.

Similar to the discussions in Hering and Genton (2011), the estimation of covariance structures given the estimates of the trend are required and indeed play an important role in the proposed test. Our simulations also show that a misspecified model can yield poor sizes and powers. In practice, it is often unrealistic to specify the true model underlying the data. If an insufficient model is naively used in the estimation, our test results may be misleading. Therefore, the estimation must respect the covariance structure of the real data to its largest extent to retain the accuracy and the power of the test. Various stationary and nonstationary covariance models have been developed to accommodate the idiosyncratic features of a given data set (see, e.g., Sampson and Guttorp, 1992; Cressie and Huang, 1999; Gneiting, 2002; Stein, 2005; Paciorek and Schervish, 2006; Gneiting *et al.*, 2007; Jun and Stein, 2007, 2008), which should be further explored to test for the most appropriate choice for improving the power of our test.

Because of the temporal correlation in the CFR, we have treated the spatial reconstruction at each different year separately. In principal, our method can be employed to compare two space–time reconstructions as a whole. In such a case, a common space–time covariance function of two CFR residual processes will first be estimated and then used together with the spatially varying mean to scale each of the two CFR residual processes. This procedure will yield two uncorrelated samples that are suitable for the two-sample KS test. The *p*-value of the KS test reflects the distance between two CFRs with their space–time correlation structure taken into account. However, this will lead to the computational difficulty that often rises for large spatial data. Indeed, this concern is also part of the reason why we compare the spatial reconstructions regionally. Because the spatial data analysis is typically vulnerable to the data size, dealing with climate data will inevitably require techniques for large-data analysis. Sun *et al.* (2012) provided a through review of the most recent advances in large-data analysis techniques.

We have only compared the best estimates from each method but have not examined the quality of uncertainty accompanied with each reconstruction. Moreover, in this paper, we have not included the Bayesian reconstruction that has risen as a capable method to integrate the physical constraints and proxies (Tingley and Huyber, 2010a, 2010b; Li *et al.*, 2010) and to provide a full account of uncertainties for correlated process. The comparisons between the traditional CFRs discussed herein and Bayesian reconstructions that incorporate explicit uncertainty estimates will be an important future research direction. Finally, if a new CFR is shown to be significantly different from those currently considered by rejecting the null hypothesis of their comparison, we would more closely investigate whether the difference is due to the mean, the variance, or the correlation structure, as well as where the difference occur. Additionally, if a CFR faithfully recovers the target, its residual process will be an approximate white noise process, and the smaller the variance of the noise process, the more precise the CFR will be. Our method can therefore be adapted to assess and rank the quality of significantly different CFRs by comparing them to the white noise process while having the magnitude of variance considered.

Spatial assessments of the CFRs confirm a recent conclusion of Smerdon *et al.* (2011a), who argued that pseudoproxy evaluations of currently applied multivariate linear regression techniques suggest only small differences in their relative performance. Our assessment metric suggests that the discrepancies between CFRs derived from the five reconstruction methods are, in general, not statistically significant when they are evaluated by comparing their underlying structures, albeit regional and temporal differences in the testing results are observed. This conclusion is not surprising given the fact that these methods were all built on the same statistical model. Nevertheless, it is still evident that TTH is mostly distinct from the other reconstruction methods despite being close to TTL, whereas CCA is mostly indistinguishable from

the others. This finding points to the hybrid spectral calibration approach as an important feature of TTH, which should be further evaluated for its impact on the spatial performance of derived CFRs.

The overall lack of statistically significant differences between the derived CFRs is a critical result given the extensive discussions in the literature about the singular performance of one method over another (e.g., Mann *et al*., 2005, 2007; Rutherford *et al*., 2010). These arguments have been based largely on assessments of NH mean skill, which are still valid. Nevertheless, the fundamental purpose of CFRs is to provide spatial information on past climate variability. Toward such end, no current multivariate linear regression method appears to provide substantial performance advantages over another. Moreover, the residual processes are characterized by dramatically varying means and nonstationary covariance structures, suggesting that the CFRs are statistically distinct from the climate target they are attempting to estimate. These facts should provide important guidance to the CE paleoclimate community in recognition of the fact that current linear statistical models may not be the ideal approach for large-scale field reconstructions of CE temperature, regardless of methodological decisions within this framework such as regularization or calibration choices.

With these results in mind, it appears that increases in proxy sampling are very important for improving large-scale CFRs, regardless of the method employed (Smerdon *et al.*, 2011a). Moreover, methodological work should focus on frameworks that incorporate physically based models as constraints on the reconstruction problem (e.g., Evans *et al*., 2006; Anchukaitis *et al*., 2006; Schmidt *et al*., 2007; Tolwinski-Ward *et al*., 2010 Thompson *et al*., 2011). Recent Bayesian studies have provided the groundwork for such approaches (Li *et al*., 2010; Tingley and Huybers, 2010a,b), and paleoclimatic assimilation techniques have also shown promise (Goosse *et al*., 2010; Widmann *et al*., 2010). Further work to fully vet multivariate linear regression techniques is of course warranted, which can also be used with additional physical constraints (e.g., Kaplan *et al*., 2003). The fundamental need to address the underlying statistical models in future reconstruction work, nevertheless, is an important emphasis moving forward, and our results further support such a focus.

Acknowledgements

Li's research was partially supported by National Science Foundation grant DMS-1007686, and Smerdon's research was partially supported by National Science Foundation grant ATM0902436 and National Oceanic and Atmospheric Administration grants NA07OAR4310060 and NA10OAR4320137. The authors thank the editor, the special issue editors, and the referees for constructive suggestions that have improved the content and presentation of this article.

REFERENCES

Anchukaitis KJ, Evans MN, Kaplan A, Vaganov EA, Hughes MK, Grissino-Mayer HD, Cane MA. 2006. Forward modeling of regional-scale tree-ring patterns in the southeastern United States and the recent emergence of summer drought stress. *Geophysical Research Letters* **33**(4): L04705, DOI: 10.1029/2005GL025050.

Anchukaitis KJ, Buckley BM, Cook ER, Cook BI, D'Arrigo RD, Ammann CM. 2010. The influence of volcanic eruptions on the climate of the Asian monsoon region. *Geophysical Research Letters* **37**: L22703, DOI: 10.1029/2010GL044843.

Ammann CM, Joos F, Schimel DS, Otto-Bliesner BL, Tomas RA. 2007. Solar influence on climate during the past millennium: Results from transient simulations with the NCAR Climate System Model. *Proceedings of the National Academy of Sciences* **104**: 3713–3718.

Atger F. 2003. Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: consequences for calibration. *Monthly Weather Review* **131**: 1509–1523.

Bliznyuk N, Carroll RJ, Genton MG, Wang Y. 2012. Variogram estimation in the presence of trend. *Statistics and Its Interface*. in press.

Briggs WM, Levine RA. 1997. Wavelets and field forecast verification. *Monthly Weather Review* **125**: 1329–1341.

Brohan P, Kennedy JJ, Harris I, Tett SFB, Jones PD. 2006. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *Journal of Geophysical Research* **111**: D12106, DOI: 10.1029/2005JD006548.

Buckley BM, Anchukaitis KJ, Penny D, Fletcher R, Cook ER, Sano M, Nam LC, Wichienkeeo A, Minh TT, Hong TM. 2010. Climate as a contributing factor in the demise of Angkor, Cambodia. *Proceedings of the National Academy of Sciences* **107**: 6748–6752.

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd Edition. Chapman & Hall: Boca Raton, FL.

Cook ER, Krusic P. 2004. *The North American Drought Atlas*. NOAA Paleoclimatology: Boulder, CO.

Cook ER, Anchukaitis KJ., Buckley BM, DArrigo RD, Jacoby GC, Wright WE. 2010. Asian monsoon failure and megadrought during the last millennium. *Science* **328**(5977): 486–489.

Cook ER, Seager R, Cane MA, Stahle DW. 2007. North American drought: Reconstructions, causes, and consequences. *Earth-Science Reviews* **81**(1–2): 93–134.

Cressie N, Huang H-C. 1999. Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association* **94**: 1330–1340.

Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**: 253–263.

Evans MN, Kaplan A, Cane MA. 2002. Pacific sea surface temperature field reconstruction from coral δ18O data using reduced space objective analysis. *Paleoceanography* **17**: 1007, DOI: 10.1029/2000PA000590.

Evans MN, Reichert BK, Kaplan A, Anchukaitis KJ, Vaganov EA, Hughes MK, Cane MA. 2006. A forward modeling approach to paleoclimatic interpretation of tree-ring data. *Journal of Geophysical Research* **111**: G03008, DOI: 10.1029/2006JG000166.

Francisco-Fernandez M, Opsomer JD. 2005. Smoothing parameter selection methods for nonparametric regression with spatially correlated errors. *The Canadian Journal of Statistics* **33**: 279–295.

Gneiting T. 2002. Nonseparable, stationary covariance functions for spaceÐtime data. *Journal of the American Statistical Association* **97**: 590–600.

Gneiting T, Genton MG, Guttorp P. 2007. Geostatistical space-time models, stationarity, separability and full Symmetry. In *Statistics of Spatio-Temporal Systems*, Vol. 107, Finkenstaedt B, Held L, Isham V (eds). Chapman & Hall/CRC Press: Boca Raton, FL; 151–175.

Golub GH, Heath M, Wahba G. 1979. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**: 215–223.

Gong X, Barnston AG, Ward NM. 2003. The effect of spatial aggregation on the skill of seasonal precipitation forecasts. *Journal of Climate* **16**: 3059–3071.

González-Rouco FJ, von Storch H, Zorita E. 2003. Deep soil temperature as proxy for surface air-temperature in a coupled model simulation of the last thousand years. *Geophysical Research Letters* **30**: 2116, DOI: 10.1029/2003GL018264.

González-Rouco FJ, Beltrami H, Zorita E, von Storch H. 2006. Simulation and inversion of borehole temperature profiles in surrogate climates: Spatial distribution and surface coupling. *Geophysical Research Letters* **33**: L01, 703, DOI: 10.1029/2005GL024,693.
González-Rouco FJ, Fernandez-Donado L, Raible CC, Barriopedro D, Luterbacher J, Jungclaus JH, Swingedouw D, Servonnat J, Zorita E, Wagner S, Ammann CM. 2011. Medieval climate anomaly to little ice age transition as simulated by current climate models. *PAGES news* **19**(1): 7–8.
Goosse H, Crespin E, de Montety A, Mann ME, Renssen H, Timmermann A. 2010. Reconstructing surface temperature changes over the past 600 years using climate model simulations with data assimilation. *Journal of Geophysical Research* **115**: D09108, DOI: 10.1029/2009JD012737.
Hering A, Genton MG. 2011. Comparing spatial predictions. *Technometrics* **53**: 414–425.
Hoerl AE, Kennard RW. 1970. Ridge regression: Applications to nonorthogonal problems. *Technometrics* **12**: 69–82.
Huang C, Hsing T, Cressie N. 2008. Nonparametric estimation of varioram and its spectrum. *Technical Report Technical Report No. 08-05*, Indiana University.
Jansen E, Overpeck J, Briffa KR, Duplessy J-C, Joos R, Masson-Delmotte V, Olago D, Otto-Bliesner B, Peltier WR, Rahmstorf S, *et al.* 2007. Palaeoclimate. 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*, Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds). Cambridge University Press: Cambridge, United Kingdom and New York, NY, USA.
Jones PD, Briffa KR, Osborn TJ, Lough JM, van Ommen TD, Vinther BM, Luterbacher J, Wahl ER, Zwiers FW, Mann ME, *et al.* 2009. High resolution paleoclimatology of the last millennium: a review of current status and future prospects. *Holocene* **19**: 3–49.
Jun M, Stein ML. 2007. An approach to producing space-time covariance functions on spheres. *Technometrics* **49**: 468–479.
Jun M, Stein ML. 2008. Nonstationary covariance models for global data. *Annals of Applied Statistics* **2**: 1271–1289.
Kaplan A, Cane MA, Kushnir Y. 2003. Reduced space approach to the optimal analysis interpolation of historical marine observations: Accomplishments, difficulties, and prospects. In *Advances in the Applications of Marine Climatology: The Dynamic Part of the WMO Guide to the Applications of Marine Climatology*. WMO/TD-1081, World Meteorological Organization: Geneva, Switzerland; 199–216.
Kotz S, Nadarajah S. 2004. *Multivariate t distributions and their applications*. Cambridge University Press: Cambridge.
Krishnaiah PR, Mudholkar GS, Subbaiah P. 1980. Simultaneous test procedures for mean vectors and covariance matrices. In *Handbook of Statistics: Analysis of Variance*, Krishnaiah PR (ed.). North-Holland: Netherlands; 631–671.
Li B, Nychka DW, Ammann CM. 2010. The value of multi-proxy reconstruction of past climate (with discussions and rejoinder). *Journal of the American Statistical Association* **105**: 883–911.
Lund R, Li B. 2009. Revisiting climate region definitions via clustering. *Journal of Climate* **22**: 1787–1800.
Mann ME, Bradley RS, Hughes MK. 1998. Global-scale temperature patterns and climate forcing over the past six centuries. *Nature* **392**: 779–787.
Mann ME, Rutherford S, Wahl E, Ammann C. 2005. Testing the fidelity of methods used in proxy-based reconstructions of past climate. *Journal of Climate* **18**: 4097–4107.
Mann ME, Rutherford S, Wahl E, Ammann C. 2007. Robustness of proxy-based climate field reconstruction methods. *Journal of Geophysical Research* **112**: D12109, DOI: 10.1029/2006JD008272.
Mann ME, Zhang Z, Hughes MK, Bradley RS, Miller SK, Rutherford S, Ni F. 2008. Proxybased reconstructions of hemispheric and global surface temperature variations over the past two millennia. *Proceedings of the National Academy of Sciences* **105**: 13252–13257.
Mann ME, Zhang Z, Rutherford S, Bradley RS, Hughes MK, Shindell D, Ammann C, Faluvegi G, Ni F. 2009a. Global signatures and dynamical origins of the little ice age and the medieval climate anomaly. *Science* **326**: 1256–1260.
Mann ME, Woodruff JD, Donnelly JP, Zhang Z. 2009b. Atlantic hurricanes and climate of the past 1,500 years. *Nature* **460**: 880–883.
Mardia KV, Kent JT, Bibby JM. 1979. *Multivariate Analysis*. Academic Press: London.
Neukom R, Luterbacher J, Villalba R, Kttel M, Frank D, Jones PD, Gosjean M, Esper J, Lopez L, Wanner H. 2010. Multi-centennial summer and winter precipitation variability in souther South America. *Geophysical Research Letters* **37**: L14708, DOI: 10.1029/2010GL043680.
North GR, Biondi F, Bloomfield P, Christy JR, Cuffey KM, Dickinson RE, Druffel ERM, Nychka D, Otto-Bliesner B, Roberts N, *et al.* 2006. *Surface Temperature Reconstructions for the Last 2,000 Years*. The National Academies Press: Washington DC.
North GR, Wang J, Genton MG. 2011. Correlation models for temperature fields. *Journal of Climate* **24**: 5850–5862.
Opsomer J, Wang Y, Yang Y. 2001. Nonparametric regression with correlated errors. *Statistical Science* **16**: 134–153.
Paciorek CJ, Schervish MJ. 2006. Spatial modelling using a new class of nonstationary covariance functions. *Environmetircs* **17**: 483–506.
Pavlicová M, Santer TJ, Cressie N. 2008. Detecting signals in FMRI data using powerful FDR procedures. *Statistics and Its Interface* **1**: 23–32.
Riedwyl N, Kuttel M, Luterbacher J, Wanner H. 2009. Comparison of climate field reconstruction techniques: Application to Europe. *Climate Dynamics* **32**: 381–395.
Rutherford S, Mann ME, Delworth TL, Stouffer RJ. 2003. Climate field reconstruction under stationary and nonstationary forcing. *Journal of Climate* **16**: 462–479.
Rutherford SD, Mann ME, Ammann CM, Wahl ER. 2010. Comments on:"a surrogate ensemble study of climate reconstruction methods: stochasticity and robustness. *Journal of Climate* **23**: 2832–2838.
Sampson PD, Guttorp P. 1992. *Nonparametric Estimation of Nonstationary Spatial Covariance Structure* **87**: 108–119.
Schmidt GA. 2010. Enhancing the relevance of paleoclimatic model/data comparisons for assessments of future climate change. *J Quaternary Sci* **25**: 79–87.
Schmidt GA, LeGrande AN, Hoffmann G. 2007. Water isotope expressions of intrinsic and forced variability in a coupled ocean-atmosphere model. *Journal of Geophysical Research* **112**: D10103, DOI: 10.1029/2006JD007781.
Shen X, Huang HC, Cressie N. 2002. Nonparametric hypothesis testing for a spatial signal. *Journal of the American Statistical Association* **97**: 1122–1140.
Smerdon JE, Gonzlez-Rouco JF, Zorita E. 2008b. Comment on "Robustness of proxy-based climate field reconstruction methods" by Michael E. Mann et al. *Journal of Geophysical Research* **113**: D18106, DOI: 10.1029/2007JD009.
Smerdon JE, Kaplan A, Chang D. 2008a. On the standardization sensitivity of RegEM climate field reconstructions. *Journal of Climate* **21**: 6710–6723.
Smerdon JE, Kaplan A, Amrhein DE. 2010. Erroneous model field representations in multiple pseudoproxy studies: Corrections and implications. *Journal of Climate* **23**: 5548–5554.
Smerdon JE. 2012. Climate models as a testbed for climate reconstruction methods: pseudoproxies. *Wiley Interdisciplinary Reviews Climate Change* **3**: 63–77.
Smerdon JE, Kaplan A, Zorita E, González-Rouco JF, Evans MN. 2011a. Spatial performance of four climate field reconstruction methods targeting the Common Era. *Geophysical Research Letters* **38**: L11705, DOI: 10.1029/2011GL047372.
Smerdon JE, Kaplan A, Chang D, Evans MN. 2011b. A pseudoproxy evaluation of the CCA and RegEM methods for reconstructing climate fields of the last millennium. *Journal of Climate* **24**: 1284–1309.
Snell SE, Gopal S, Kaufmann RK. 2000. Spatial interpolation of surface air temperatures using artificial neural networks: Evaluating their use for downscaling GCMs. *Journal of Climate* **13**: 886–895.
Stein M. 1999. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag: New York.
Stein ML. 2005. Space-time Covariance Functions. *Journal of the American Statistical Association* **100**: 310–321.
Stein ML. 2011. Editorial. *Annals of Applied Statistics* **5**(1): 1–4.
Sun Y, Li B, Genton MG. 2012. Geostatistics for large datasets. In *Space-Time Processes and Challenges Related to Environmental Problems*, Vol. 207, Chapter 3, Porcu E, Montero JM, Schlather M (eds). Springer: Berlin, London; 55–77.

Tingley MP, Huybers PA. 2010a. Bayesian algorithm for reconstructing climate anomalies in space and time. Part I: Development and applications to paleoclimate reconstruction problems. *Journal of Climate* **23**: 2759–2781.

Tingley MP, Huybers PA. 2010b. Bayesian algorithm for reconstructing climate anomalies in space and time. Part II: Comparison with the regularized expectation maximization algorithm. *Journal of Climate* **23**: 2782–2800.

Tingley MP, Craigmile PF, Haran M, Li B, Mannshardt-Shamseldin E, Rajaratnam B. 2012. Piecing together the past: Statistical insights into paleoclimatic reconstructions. *Quaternary Science Reviews*. in press.

Thompson DM, Ault TR, Evans MN, Cole JE, Emile-Geay J. 2011. Comparison of observed and simulated tropical trends using a forward model of coral d18O. *Geophysical Research Letters* **38**: L14706, DOI: 10.1029/2011GL048224.

Tolwinski-Ward SE, Evans MN, Hughes MK, Anchukaitis KJ. 2010. An efficient forward model of the climate controls on interannual variation in tree-ring width. *Climate Dynamics* **36**: 2419–2439, DOI: 10.1007/s00382-010-0945-5.

von Storch H, Zorita E, Jones JM, Dimitriev Y, Gonzlez-Rouco F, Tett SFB. 2004. Reconstructing past climate from noisy data. *Science* **306**: 679–682.

von Storch H, Zorita E, Jones JM, Dimitriev Y, Gonzlez-Rouco F, Tett SFB. 2006. Response to comment on "Reconstructing past climate from noisy data". *Science* **312**: 529.

Wang W, Anderson BT, Entekhabi D, Huang D, Su Y, Kaufmann RK, Potter C, Myneni RB. 2007. Intraseasonal interactions between temperature and vegetation over the boreal forests. *Earth Interactions* **11**: 1–30.

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

Willis HL. 2002. *Spatial Electric Load Forecasting*. Marcel-Dekker, Inc: New York, NY.

Zhang Z, Mann ME, Cook ER. 2004. Alternative methods of proxy-based climate field reconstruction: Application to summer drought over the conterminous United States back to AD 1700 from tree-ring data. *Holocene* **14**: 502–516.

