Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Evaluating Proxy Influence in Assimilated Paleoclimate Reconstructions—Testing the Exchangeability of Two Ensembles of Spatial Processes

Trevor Harris, Bo Li, Nathan J. Steiger, Jason E. Smerdon, Naveen Narisetty & J. Derek Tucker

To cite this article: Trevor Harris, Bo Li, Nathan J. Steiger, Jason E. Smerdon, Naveen Narisetty & J. Derek Tucker (2020): Evaluating Proxy Influence in Assimilated Paleoclimate Reconstructions—Testing the Exchangeability of Two Ensembles of Spatial Processes, Journal of the American Statistical Association, DOI: 10.1080/01621459.2020.1799810

To link to this article: https://doi.org/10.1080/01621459.2020.1799810

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 00, NO. 0, 1–14: Applications and Case Studies
https://doi.org/10.1080/01621459.2020.1799810

Evaluating Proxy Influence in Assimilated Paleoclimate Reconstructions—Testing the Exchangeability of Two Ensembles of Spatial Processes

Trevor Harris[a], Bo Li[a], Nathan J. Steiger[b], Jason E. Smerdon[b], Naveen Narisetty[a], and J. Derek Tucker[c]

[a]Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL; [b]Department of Oceanography, Lamont-Doherty Earth Observatory, Palisades, NY; [c]Department of Statistical Sciences, Sandia National Laboratories, Albuquerque, NM

ABSTRACT

Climate field reconstructions (CFRs) attempt to estimate spatiotemporal fields of climate variables in the past using climate proxies such as tree rings, ice cores, and corals. Data assimilation (DA) methods are a recent and promising new means of deriving CFRs that optimally fuse climate proxies with climate model output. Despite the growing application of DA-based CFRs, little is understood about how much the assimilated proxies change the statistical properties of the climate model data. To address this question, we propose a robust and computationally efficient method, based on functional data depth, to evaluate differences in the distributions of two spatiotemporal processes. We apply our test to study global and regional proxy influence in DA-based CFRs by comparing the background and analysis states, which are treated as two samples of spatiotemporal fields. We find that the analysis states are significantly altered from the climate-model-based background states due to the assimilation of proxies. Moreover, the difference between the analysis and background states increases with the number of proxies, even in regions far beyond proxy collection sites. Our approach allows us to characterize the added value of proxies, indicating where and when the analysis states are distinct from the background states. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2019
Accepted July 2020

KEYWORDS
Climate field reconstructions; Data assimilation; Functional depth; Spatial fields

1. Introduction

Since their first high-profile application two decades ago (Mann, Bradley, and Hughes 1998), multi-proxy spatiotemporal climate field reconstructions (CFRs) have become increasingly popular in the climate science community for their ability to reconstruct global climate variability on seasonal and annual timescales over many hundreds of years into the past (Jones et al. 2009; Smerdon and Pollack 2016; Christiansen and Ljungqvist 2017). The reconstruction of climate is critical because data from instrumental observations are only available for the past 100–150 years. CFRs therefore provide estimates of past climate variability and extreme events that may not be well represented over the instrumental interval. This helps to better characterize the physical dynamics of the climate system and how climate may change in the future.

The basic approach of CFRs is to statistically relate a collection of climate proxies, such as isotopic information in ice cores, the width of tree rings, or coral isotope data, to observed climate variables like temperature and soil moisture during their periods of overlap (Masson-Delmotte et al. 2013). Once the relationship between the proxies and the climate variables is established, the proxies are used to estimate climate variability during periods when observations are not available in the past. CFRs thus depend critically on the imperfect proxy information and the robustness with which their relationship to observed climate variables can be defined. A central approach to this problem in the past has been through regularized versions of multivariate regression techniques (e.g., Lee, Zwiers, and Tsao 2008; Jones et al. 2009; Smerdon 2012; Tingley et al. 2012; Guillot, Rajaratnam, and Emile-Geay 2015; Li, Zhang, and Smerdon 2016; Smerdon and Pollack 2016; Christiansen and Ljungqvist 2017). More advanced techniques have been emerging, however, all of which are associated with advantages and challenges that require further evaluation and assessment.

A recent CFR innovation is the paleoclimatic data assimilation (DA) algorithms, which are a class of reconstruction methods that optimally combine general circulation models (GCMs) with proxy information to create paleoclimate reconstructions (Goosse et al. 2012; Steiger et al. 2014; Hakim et al. 2016; Steiger et al. 2018; Tardif et al. 2019). The primary advantage of DA approaches is their ability to jointly reconstruct multiple atmosphere-ocean variables and to do so in a manner that is physically consistent within the framework of a climate model. An important distinction of the DA methods, relative to the other statistical (inverse) methods, is the use of forward models that map from climate states to the proxies. An additional advantage is that DA algorithms naturally provide probabilistic, ensemble estimates of past climate. Such ensemble reconstructions first begin with a background ensemble of states from a climate model. These states are then updated through the equations of DA (Steiger

CONTACT Trevor Harris trevorh2@illinois.edu Department of Statistics, University of Illinois at Urbana-Champaign, 725 S Wright St, Champaign, IL 61820.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

et al. 2014), based on the available proxy information and the uncertainties involved, to arrive at an analysis ensemble state estimate. This probabilistic analysis state provides an uncertainty quantification that is critical given the noisy relationship between paleoclimate proxies and climate variables.

Despite the rapid development of DA-based reconstruction methods, much remains to be characterized about the influence of each of their two components: climate models and paleoclimate proxies. In currently published DA-based CFRs (e.g., Steiger et al. 2018), it is hard to quantify how much information the models and the proxies each contribute to the end product. One approach is independent proxy validation of the analysis states (Hakim et al. 2016). Another approach is to compare climate time series and climate patterns in the background and analysis with each other and with observations (Singh et al. 2018). However, more formal statistical approaches are called for to differentiate whether or not the climate model-based background is fundamentally distinct from the analysis. If the background and analysis are not in fact distinct, then this would imply that DA-based CFRs are essentially dominated by the underlying climate model and fail to glean information from the historical proxy data. A lack of proxy influence would, therefore, indicate a need to fundamentally re-evaluate DA methodologies.

In this article, we quantify the level of proxy influence in the analysis states of a DA product by introducing a robust and computationally efficient method for evaluating the exchangeability of two ensembles of random fields. The purpose of this study is therefore 2-fold: to answer an important climatological question by quantifying and assessing the influence of proxies in a new DA based CFR product and to develop a new statistical test for comparing the distributions of two sets of random fields. In the following two subsections, we provide background on the methodological development embodied in this article and the characteristics of the DA-based CFR that we analyze.

1.1. Previous Work in Random Fields Comparisons

Comparing two spatial processes has been addressed in both the geostatistics and functional data analysis literature. The general strategy in both frameworks is to reduce the dimension of the random process either by a low-rank decomposition or by parameterization and then to develop a test for evaluating differences in the reduced dimension.

The wavelet decomposition has been widely used to reduce a stochastic process to a finite number of wavelet coefficients, then the comparison between two processes can be transformed into the comparison between two sets of wavelet coefficients (Briggs and Levine 1997; Shen, Huang, and Cressie 2002; Pavlicova, Santer, and Cressie 2008). Snell, Gopal, and Kaufmann (2000) and Wang et al. (2007) introduced methods for comparing random fields based on their spatial interpolation root-mean-square error and R^2 coefficient. Their methods were later extended by Hering and Genton (2011) to include more arbitrary loss functions. Motivated by Lund and Li (2009) that compared two time series, Li and Smerdon (2012) proposed a parametric method to jointly assess the first two moments between two random fields.

Functional data analysis approaches assume that the spatial random fields are noisy realizations of an underlying continuous function. The majority of existing functional approaches have focused on testing the equality of the mean functions arising from two functional datasets (Ramsay and Silverman 2005; Zhang and Chen 2007; Horvath, Kokoszka, and Reeder 2013; Staicu et al. 2014), although more recently the second-order structure of functional data has also been considered (Zhang and Shao 2015). Li, Zhang, and Smerdon (2016) extended Zhang and Shao (2015) to evaluate the joint difference in mean and covariance structure as well as in the trend surface between two spatiotemporal random fields. A nice feature of functional data analysis methods, as opposed to geostatistical methods, is that assumptions about distribution and model specification can be relaxed if there are replicate observations in the data.

All the above procedures are nevertheless inadequate for our problem because the proxies can simultaneously affect the mean, covariance, and higher order structures of the reconstructed climate field. The most comprehensive way to identify proxy influence is therefore to compare the distributions of the background and analysis states. The rich ensemble structure of the background and analysis states also allows us to examine more information than differences in the mean and covariance parameters. We take advantage of the ensembles by employing a functional data approach that is both distribution and parameter free.

The problem of comparing the distributions of functions has remained relatively unexplored. Hall and Van Keilegom (2007) proposed a Cramer–von Mises-like test by constructing an empirical distribution over each of the samples and measuring the $\mathbb{L}^2$ distance between the empirical distributions. Benko, Hardle, and Kneip (2009) introduced a permutation test on the leading coefficients of the common functional principal components (FPCs) and Corain et al. (2014) introduced three omnibus tests for combining pointwise tests on the observations of the functions. Each of these methods depends on a resampling procedure that renders them computationally prohibitive for large ensembles like the DA ensemble output that we consider.

Pomann, Staicu, and Ghosh (2016) proposed a method based on marginal FPCs that does not require resampling, called the functional Anderson–Darling (FAD) test. The FAD test compares the distributions of the marginal FPCs using the two sample Anderson–Darling test and a Bonferroni correction. Lopez-Pintado and Romo (2009) proposed a rank based band depth test (BAND). The BAND test is closely related to the multivariate distribution test based on the quality index (QI, Liu and Singh 1993) but it replaces the multivariate simplicial depth (Liu 1990) with the functional band depth (Lopez-Pintado and Romo 2009). Both of these tests are inadequate for our data because, as we show in the supplementary materials, they are incapable of detecting heterogeneous variance changes across the domain of the generating process. This causes the BAND test to experience a severe loss of power and for FAD to miss an important trend (Section 4.1) in our dataset.

In this article, we propose a new nonparametric statistic, based on the concept of data depth, for assessing the equality of distributions between two spatial datasets. Our test falls into

the general category of functional data analysis methods for comparing spatial random fields, but is conceptually different from previous efforts in this area. The use of data depth for comparing two multivariate distributions was first explored by Liu and Singh (1993) who introduced the QI for comparing two multivariate distributions. The QI essentially measures the mean outlyingness of one sample from a reference sample using data depth. We will extend their ideas to the functional setting and propose a modification that makes our test statistic invariant to the reference distribution. The use of depth, and particularly integrated Tukey depth (Cuevas and Fraiman 2009), ensures our test is computationally efficient, distribution free, and invariant to location, scale, warping, and other nuisance properties that could influence the testing (Nagy et al. 2016).

1.2. Reconstruction Data

The DA-based CFR that we analyze comes from the Paleo Hydrodynamics Data Assimilation product (PHYDA), which is a global paleoclimate reconstruction of both temperature and moisture variables (Steiger et al. 2018). PHYDA incorporates a simulation from the Community Earth System Model (CESM) last millennium ensemble experiment, run over the historical years 850 CE to 1850 CE (Otto-Bliesner et al. 2016).

A collection of modeled climate fields from the CESM simulation are used to form the background state in the DA scheme. For the purpose of our analysis herein, we will specifically use the modeled and reconstructed 2-m surface temperature fields. The temperature fields are processed from the native model output by annual averages and spatially discretizing onto a 2° latitude and longitude grid (144 × 96 grid points). Annual in this context is defined as the interval between April and March of the following year, thus yielding 998 such climatological years to be used for the background ensemble. Because of the large data files produced by PHYDA, we only used a 100 member sub-ensemble, randomly drawn from the original 998 member ensemble, for our analyses. The final processed background state, therefore, consists of 100 spatial fields, each observed on the same 144 × 96 grid points.

The 998 analysis states are derived from the background state by using DA to incorporate temporally available proxy information during each year of reconstruction (see Figure 1). Each analysis state is also a 100 member ensemble of 2 m surface temperature fields discretized to the same 2° latitude and longitude grid as the background ensemble. Quantifying the influence that proxies have in the analysis states is quite challenging due to their small individual effect sizes, and the fact that they can affect higher order structures of the data beyond the mean and variance. Identifying the full effect of the proxies would, therefore, require testing for distributional changes.

2. Statistical Solution

We first formulate our scientific problem into a sequence of hypothesis tests. We then introduce the integrated Tukey depth and propose our test statistic, followed by a discussion of the asymptotic distribution of our test statistics under the null hypothesis.

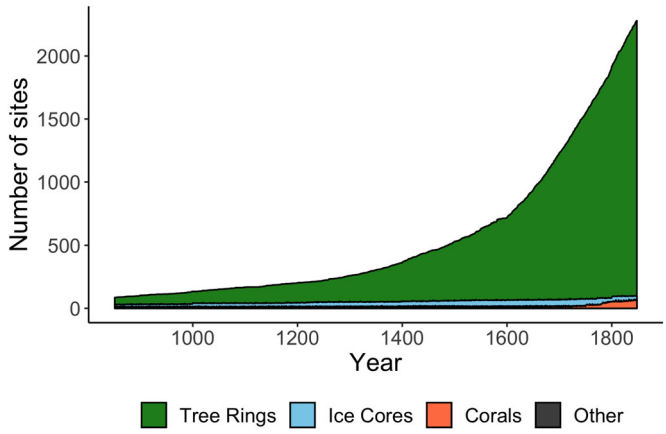


Figure 1. Temporal distribution of proxies by the three largest categories (tree rings, ice cores, and corals). There are total of 2468 proxies used over this interval, with the vast majority being tree rings.

2.1. Formulation of Evaluating Proxy Influence

Let X and Y_t, respectively, represent the ensemble in the background state and the ensemble in the analysis state at time t in PHYDA. Under the assimilation design, the proxies at time t are the only contributors to the differences between the two sets of ensembles. Our goal is to define and quantify the differences between X and Y_t each year to assess the proxy influence.

The amount that proxies impact the analysis states depends on many factors including the proxy type (e.g., tree ring, ice core, coral), where proxies were collected, and the interval over which the proxies were observed (Steiger et al. 2018). As shown in Figure 2, the effects from proxies may be small and thinly diffused over a noncontiguous area due to spatial correlations and teleconnections. In fact, most of the induced mean differences generally fall within the natural variation of the background fields. The most comprehensive approach to test for the proxy's cumulative influence is, therefore, to test for changes in the distributions of X and Y_t. We thus formulate our problem into the following hypotheses:

$$H_0 : X \overset{D}{=} Y_t \quad \text{versus} \quad H_A : X \overset{D}{\neq} Y_t, \tag{1}$$

where $\overset{D}{=}$ means equality in distribution. In addition to the outcome of these hypothesis tests at each time t, we are also equally interested in the pattern of those outcomes as t increases. Over time, the amount of proxy information available for reconstruction increases, while the background ensemble stays the same. We therefore might expect that the divergence between the background and the analysis distributions will increase over time accordingly if the proxies are having their due influence.

Under the functional data analysis regime, we assume that the observed data are generated from continuous functions combined with additive noise, instead of from a spatially correlated stochastic processes. In this framework, each ensemble member represents a single observation over a spatial domain where 144 × 96 grid points are embedded. This distinction allows us to consider each ensemble member as an iid realization of a stochastic process in a functional space.

We develop our test statistic for the testing problem (1) at any given t in a general context. For ease of notation, we suppress t

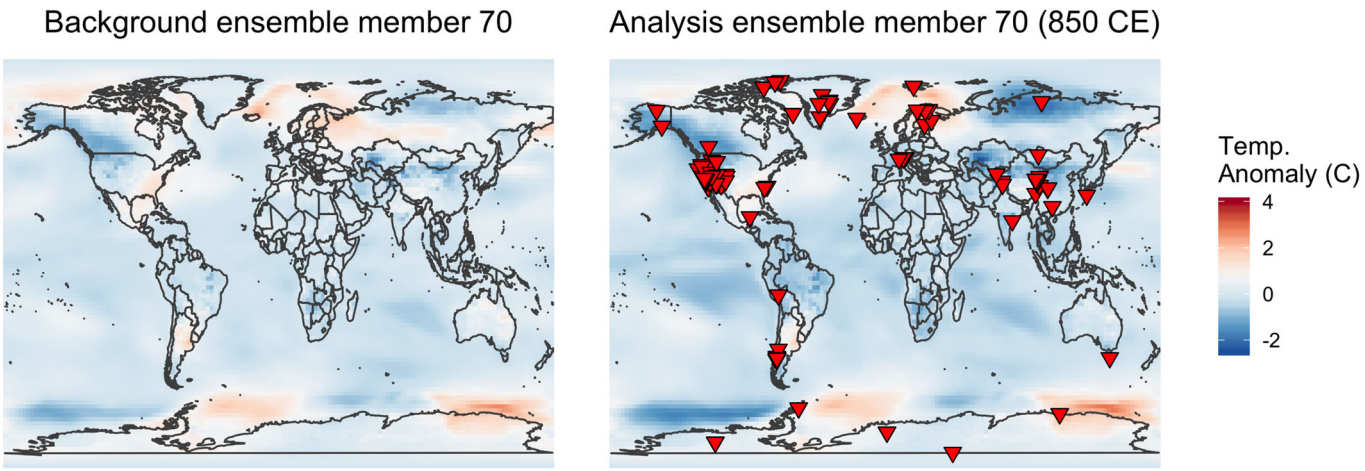


Figure 2. Temperature anomalies for a single background and analysis ensemble member with respect to the background mean field. Left panel is from the CESM simulation run while the right panel is from PHYDA during 850 CE. Red triangles indicate the locations of proxies available in 850 CE.

from Y_t. Let $X = \{X_i(s)\}_{i=1}^n$ and $Y = \{Y_j(s)\}_{j=1}^m$, where $s \in D$ and D is a compact subspace of $\mathbb{R}^p$. Without loss of generality, let D be $[0,1]^p$ and let each functional datum be observed at the same locations in $[0,1]^p$. We assume that each function X_i and Y_j is a univariate continuous function on the domain $[0,1]^p$, that is, $X_i : [0,1]^p \mapsto \mathbb{R}$ for $i \in 1,\ldots,n$; $Y_j : [0,1]^p \mapsto \mathbb{R}$ for $j \in 1,\ldots,m$. In other words, each X_i (or Y_j) is an element of the class of univariate continuous functions on $[0,1]^p$, denoted by $C[0,1]^p$. Specific to our data, we have $p = 2$ and $X_i(s)$ and $Y_j(s)$, respectively, represent the ith background state and the jth analysis state at location s.

Let P and Q be two absolutely continuous distributions on $C[0,1]^2$ and suppose each $X_i \sim P$ and each $Y_j \sim Q$. We are interested in testing if the functional data in X and in Y follow the same distribution, so (1) is equivalent to the hypotheses,

$$H_0 : P = Q; \quad \text{versus } H_A : P \neq Q, \tag{2}$$

for any given t. We will use functional data depth to construct a two sample Kolmogorov–Smirnov-type test. Other distribution free tests such as the Anderson–Darling or Cramer–von Mises test could equally have been applied. We chose Kolmogorov–Smirnov for its convenient asymptotic form and its ubiquity in testing distributions.

2.2. Integrated Tukey Depth

Data depth is a statistical concept for quantifying the centrality or "depth" of the observed data points with respect to a reference distribution. The closer an observation is to the center of the distribution, the higher its depth value should be to indicate its centrality. As the reference distribution is typically unknown, the depth of an observation has to be estimated via an empirical notion of data depth. Many notions of data depth for functional data have been developed including the integrated band depth (Lopez-Pintado and Romo 2009), extremal depth (Narisetty and Nair 2017), and various integrated univariate depths (Fraiman and Muniz 2001). Each of these depth functions has its own strengths and weaknesses but none dominates the others in all aspects, see Cuevas and Fraiman (2009) and Nagy et al. (2016) for a review. We chose the integrated Tukey depth as the basis of our test for its simplicity, robustness, computational tractability, and highly desirable theoretical properties.

Integrated depths are a well-studied class of functional data depth measures that were first introduced by Fraiman and Muniz (2001) and then studied extensively by Cuevas and Fraiman (2009) and Nagy et al. (2016). To define an integrated depth function, a univariate depth function is first defined over a collection of one-dimensional "projections" of the data which often refers to the observed values of the functions at each location $s \in D$. The univariate depth is then integrated over these projections to yield the integrated depth. Among all the univariate depths, the Tukey depth and the simplicial depth are perhaps the two most popular ones. We opted to use the Tukey depth but the simplicial depth would have been equally effective because the orderings they induce are nearly identical.

The integrated Tukey depth is defined as follows. Let P be a distribution for $X \in C[0,1]^p$, and let P_s be the marginal distribution of P at $s \in [0,1]^p$. The univariate Tukey depth of $X(s) = x(s)$ with respect to P_s is

$$D(x(s), P_s) = 1 - |1 - 2P_s(x(s))|,$$

and the integrated Tukey depth of $X = x$ with respect to P is

$$D(x, P) = \int_{[0,1]^p} D(x(s), P_s) ds.$$

To ensure that this depth function is proper, we refer to the criteria proposed by Zuo and Serfling (2000) and Mosler and Polyakova (2012). In Nagy et al. (2016), it was shown that the integrated Tukey depth satisfies translation invariance, function scale invariance, measure-preserving rearrangement invariance, maximality at the center, continuity, and quasi-concavity of the induced level sets. They also demonstrated strong universal consistency and weak uniform consistency, which assure that the integrated Tukey depth behaves well and asymptotically converges to its population counterpart under regularity conditions.

2.3. Test Statistic

We propose a test statistic KD(X, Y), called the Kolmogorov depth (KD) statistic, for our hypothesis testing problem (2) based on the integrated Tukey depth. The KD statistic measures the outlyingness of a sample $X \sim P$ from the distribution Q as well as the outlyingness of a sample $Y \sim Q$ from the distribution P. It takes the maximum of the two outlyingness measures as its value. This way we can correctly detect differences between P and Q even when they do not appear mutually outlying from each other under data depth. For example, if one of the distributions is nested inside the other then the nested distribution will not appear outlying to the other distribution.

Denote P_n as the empirical estimate of P based on the sample $X = \{X_1 = x_1, \ldots, X_n = x_n\}$ and Q_m the empirical estimate of Q based on $Y = \{Y_1 = y_1, \ldots, Y_m = y_m\}$. We start by considering P_n fixed and aim to measure the outlyingness of Q_m over P_n. To do this we first define the following two empirical measures for any given $x_k \in X$:

$$\widehat{F}_X(x_k) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(D(x_i, P_n) \le D(x_k, P_n)), \tag{3}$$

$$\widehat{G}_Y(x_k) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}(D(y_j, P_n) \le D(x_k, P_n)). \tag{4}$$

Essentially, $\widehat{F}_X(x_k)$ is a rescaling of $D(x_k, P_n)$ to its standardized rank, that is, $\widehat{F}_X(x_k) = 1/n$ if $D(x_k, P_n)$ is the smallest, $\widehat{F}_X(x_k) = 2/n$ if $D(x_k, P_n)$ is the second smallest, and so on. It acts as the empirical cumulative distribution function of $D(x_k, P_n)$ for $k \in 1, \ldots, n$ evaluated at itself and thus follows a discrete uniform distribution. The second quantity $\widehat{G}_Y(x_k)$ can be considered as the empirical cumulative distribution function of $D(y_k, P_n)$ evaluated at $D(x_k, P_n)$. Under H_0 in (2), $\widehat{G}_Y$ should be approximately uniform, so a deviation of $\widehat{G}_Y$ from the uniform distribution indicates an outlyingness of Q_m from P_n. The introduction of $\widehat{F}_X(x_k)$ and $\widehat{G}_Y(x_k)$ allows us to reduce the problem of comparing two sets of random fields to assessing the difference in distribution between two sets of random variables, $\widehat{F}_X(x_k)$ and $\widehat{G}_Y(x_k)$ for $k = 1, \ldots, n$. The latter can be naturally quantified using the Kolmogorov distance over the set X:

$$K_{P_n}(X, Y) = \max_{x_k \in X} |\widehat{F}_X(x_k) - \widehat{G}_Y(x_k)|. \tag{5}$$

To measure the outlyingness of P_n over Q_m we now fix Q_m rather than P_n. Following the same scheme, we define the two empirical measures for any given $y_k \in Y$ as

$$\widetilde{F}_X(y_k) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(D(x_i, Q_m) \le D(y_k, Q_m)),$$

$$\widetilde{G}_Y(y_k) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}(D(y_j, Q_m) \le D(y_k, Q_m)).$$

These two quantities exactly mirror $\widehat{F}_X$ and $\widehat{G}_Y$ except that now $\widetilde{G}_Y$ is uniform on the depth values $D(y_k, Q_m)$, for $k \in 1, \ldots, m$, and $\widetilde{F}_X$ is the indicator for the outlyingness of P_n from Q_m. We again take the Kolmogorov distance, but now over the set Y, as the measure of outlyingness

$$K_{Q_m}(X, Y) = \max_{y_k \in Y} |\widetilde{F}_X(y_k) - \widetilde{G}_Y(y_k)|.$$

We define the overall test statistic KD by taking the maximum of the two distances:

$$\mathrm{KD}(X, Y) = \max\{K_{P_n}(X, Y), K_{Q_m}(X, Y)\}. \tag{6}$$

The test statistic KD attains a level of symmetry by making the test invariant to the reference distribution. It is strictly nonnegative and it equals 0 only under H_0 in the hypothesis (2). Thus, the originally stated hypothesis (1) can be tested by evaluating whether KD is significantly greater than 0.

One major difference between our test statistic KD and the QI in Liu and Singh (1993) is that our test does not depend on a reference distribution while QI requires one of the samples to be used as the reference. Our test computes the outlyingness of two samples from each other and aggregates the results into one single test. This is a more efficient use of the two samples and enables KD to detect a larger range of alternative hypotheses, such as the nesting situation mentioned above. We discuss the critical values of KD in the following section.

2.4. Computing Critical Values

Deriving the asymptotic distribution of KD is nontrivial because KD explicitly depends on two non-iid processes, $D(x_i, P_n)$ and $D(y_j, Q_m)$. This renders standard results on the Kolmogorov–Smirnov test inapplicable. Nevertheless, we conjecture without formal proof that KD either follows the same limiting distribution as the regular Kolmogorov–Smirnov two sample statistic, that is,

$$\sqrt{\frac{nm}{n+m}}\mathrm{KD} \xrightarrow{D} K',$$

where

$$P(K' < t) = 1 - 2\sum_{j=1}^{\infty}(-1)^j e^{-2j^2t^2},$$

or converges to a distribution that can be closely approximated by K'. Although we are unable to prove this result in its full generality, we consider two special cases below and show that both conform to the conjecture of converging to K'. Our extensive simulation studies in Section 3 demonstrate convergence in the general case.

We first consider a special case where P is known and we are interested in testing if $Y_j \sim P$ for $j = 1, \ldots, m$. In this case, $\widehat{F}_X(x_k)$ in (3) becomes the uniform$[0, 1]$ distribution at $D(x_k, P) \in [0, 1]$. Then $K_{P_n}(X, Y)$ in (5), which is the test statistic in this special case, reduces to

$$K_P(Y) = \sup_{x_k} |D(x_k, P) - \widehat{G}_Y(x_k)|.$$

Because $\widehat{G}_Y(x_k)$ is an empirical distribution of the iid random variables $\{D(y_1, P), \ldots, D(y_m, P)\}$ at $D(x_k, P)$, $K_P(Y)$ is exactly the one sample Kolmogorov–Smirnov statistic for testing the uniformity of $\widehat{G}_Y(x_k)$. Therefore,

$$\sqrt{m}K_P(Y) \xrightarrow{D} K'.$$

We further consider another special case where P and Q are both unknown but with either $n \gg m$ or $m \gg n$. We can show that $K_{P_n}(X, Y)$ (or $K_{Q_m}(X, Y)$) converges to the Kolmogorov distribution under $n \gg m$ ($m \gg n$). We encapsulate this result in the following proposition.

Proposition 2.1. Suppose that $n \gg m$, then under the null hypothesis,

$$\sqrt{\frac{nm}{n+m}} K_{P_n}(X, Y) \xrightarrow{D} K',$$

where K' follows the Kolmogorov distribution.

The proof is deferred to the Appendix.

Generalizing the results of these special cases is challenging. This issue was also noted in Liu and Singh (1993) where the authors conjectured that their two sample QI asymptotically followed a normal distribution, as its one sample version does. Their conjecture was only later proven in Zuo and He (2006) after substantial theoretical development. The techniques that emerged from the proof in Zuo and He (2006) relied heavily on QI being an expectation, making them largely inapplicable to our context involving suprema. Proving the conjecture would require the development of advanced theoretical machinery that can accommodate the complex dependence nature of the distribution functions of the depth measures. We leave this problem open for independent theoretical research in the future.

In lieu of the proposed asymptotic distribution, we may consider using permutations to find critical values for KD (Good 2013). Permutation works well for small samples or sparsely observed functions, but it quickly becomes computationally infeasible on large volumes of data, such as our reconstruction data. For this reason, the conjectured Kolmogorov distribution is more appealing in practice.

3. Simulation Study

Simulation studies are conducted to assess the convergence of KD to K', and the size and power of the test. Each of these properties is evaluated using two-dimensional functional data because our main application considers ensembles of spatial fields. All functional data in the simulation are generated from Gaussian random processes with the Matérn covariance function (Stein 2012),

$$C(x, x') = \frac{\sigma\sqrt{\pi} r^{2\nu}}{2^{\nu-1}\Gamma(\nu + 1/2)} \left(\frac{\|x - x'\|}{r}\right)^{\nu} K_{\nu}\left(\frac{\|x - x'\|}{r}\right),$$

where Γ is the Gamma function, K_{ν} is a modified Bessel function, σ is the marginal variance of the random process, and r and ν are two nonnegative parameters called range and smoothness. The range parameter, r, governs how quickly the correlation decays between points. The smoothness parameter, ν, determines how smooth sampled functions are in terms of their differentiability.

In each simulation, we consider the sample X as the baseline and Y as the sample to be varied. For the size and convergence simulations, the marginal variance σ will always be set to 1, while r and ν will be allowed to vary. For the power simulations, μ, σ, r, and ν will all be allowed to vary.

3.1. Convergence

We use simulations to validate the conjectured asymptotic Kolmogorov distribution of our test statistic (6) under the null hypothesis. The main idea is to evaluate how well the permutation distribution of the test statistic is approximated by the Kolmogorov distribution, even at moderate sample sizes. Functional data $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_n\}$ are each generated with mean, $\mu = 0$, and standard deviation, $\sigma = 1$, on the spatial domain $[0, 1] \times [0, 1]$. Because the integrated Tukey depth is invariant to the location and scale of functional data, we only vary the range and smoothness of the covariance function: 0.2, 0.3, 0.4, 0.5 and 0.5, 1.0, 1.5, respectively. The number of replicates, n, in each sample is also varied between 25, 50, 75, 100. We also considered the unbalanced sample size case by fixing the number of replicates in Y to be 75 and allowing the number in X to vary between 25, 50, 75, 100. The results were nearly identical as to those when the sample sizes were balanced so only the balanced case is presented here.

The permutation distribution was constructed by recomputing KD on 500 permutations of the generated X and Y samples. We then calculated the $\mathbb{L}^2$ distance between the permutation distribution and the Kolmogorov distribution and the difference between critical values derived from the permutation and Kolmogorov distribution at three common significance levels: 0.01, 0.05, and 0.10. Due to the computational cost of constructing permutation distributions, we ran 100 simulations for each combination of r, ν, and n to obtain the boxplots in Figures 3 and 4.

Figure 3 demonstrates convergence of the permutation distribution to Kolmogorov in $\mathbb{L}^2$. For even small sample sizes, such as $n = 25$, the distance between the two distributions is already vanishingly small for smooth data ($r \geq 0.3$ and $\nu \geq 1.0$). The largest deviations are only observed when both the range and smoothness are small, $r < 0.3$ and $\nu < 1.0$. This is typically not an issue in practice because functional data are generally preprocessed with a smoothing step; effectively increasing ν and r. In all cases the $\mathbb{L}^2$ norm decreases rapidly with an increasing sample size such that the convergence even applies to unprocessed noisy data if the sample sizes are large enough.

Figure 4 evaluates the convergence of the two sets of critical values at the significance levels 0.01, 0.05, and 0.10. This figure shows that testing decisions reached under the asymptotic Kolmogorov distribution are generally not biased away from decisions reached under the permutation decision. Again, a sufficient amount of smoothness ($r \geq 0.3$ or $\nu \geq 1.0$) is required to have well-behaved critical values. If the data are not sufficiently smooth then the Kolmogorov distribution tends to have smaller critical values than the corresponding permutation distribution. The size will therefore be slightly inflated by using Kolmogorov and so the permutation distribution should be preferred when computationally feasible. Once a sufficient level of smoothness has been reached, in this case $r \geq 0.3$ or $\nu \geq 1.0$, the critical values of the permutation distribution become highly comparable with the Kolmogorov distribution. The observed differences are minuscule such that any decision reached using the Kolmogorov distribution is likely to be the same as if the permutation distribution were used. With noisy raw data a sufficient number of samples ($n \geq 100$) can begin to compensate for a lack of smoothness or correlation.

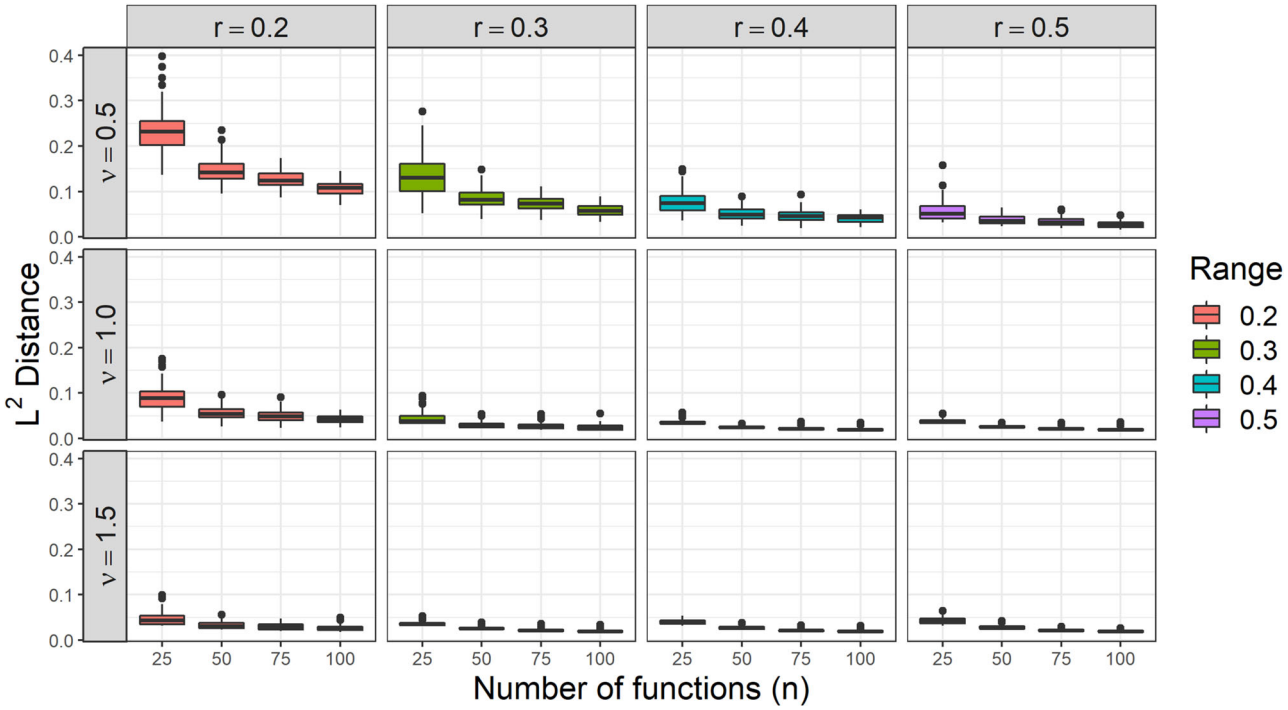


Figure 3. $\mathbb{L}^2$ distance between the permutation distribution and the Kolmogorov distribution under 12 different ranges, r, and smoothness, ν, settings.

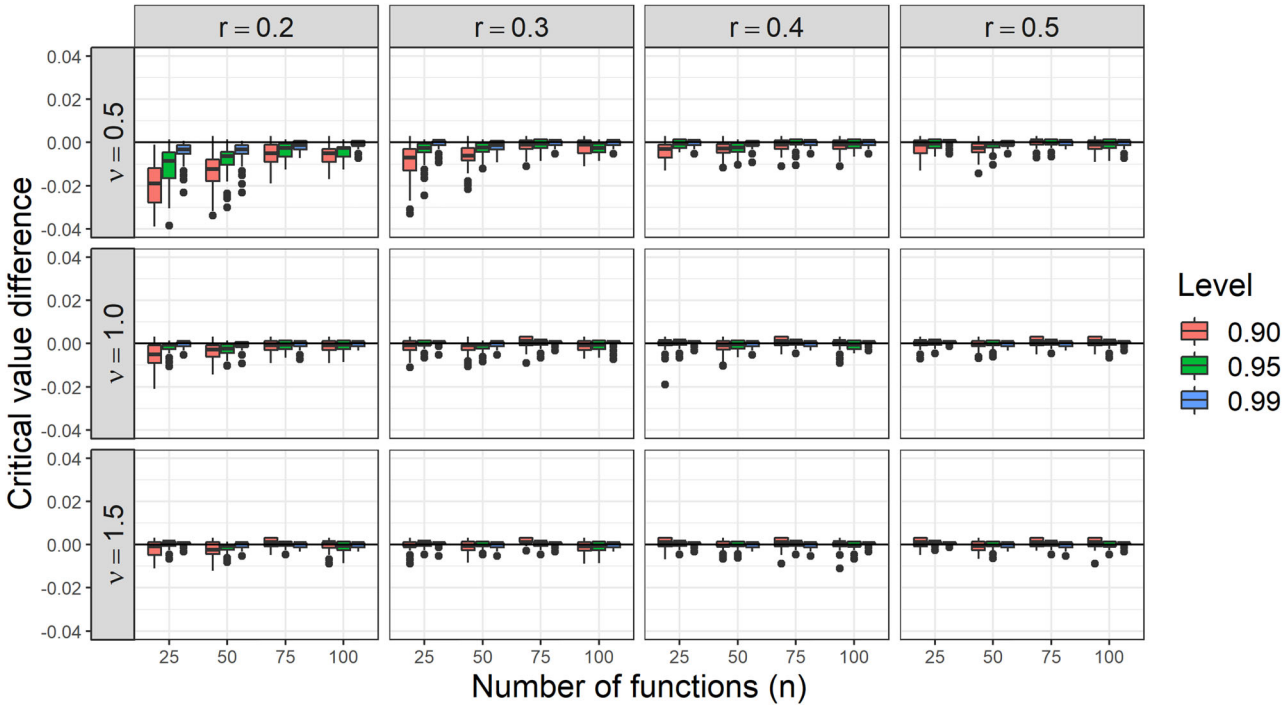


Figure 4. Kolmogorov critical values minus permutation critical values at three common test levels: 0.90, 0.95. 0.99 under 12 different ranges, r, and smoothness, ν, settings.

3.2. Size and Power

Using the same data generating process as in Section 3.1, we evaluate the size of our test using critical values from the asymptotic Kolmogorov distribution and compare our size to the QI test. Again only r, ν, and the two sample sizes n and m will be varied. The size under each combination of r, ν, n, and m was estimated using 2000 simulations; the results of which are presented in Table 1.

Our simulations show that for even small samples, such as $n = m = 50$, our test can control the size near the prescribed level if the range or smoothness is sufficiently high; that is $r \geq 0.4$ or $\nu \geq 1$. Smoothness and range are in fact more important for controlling size than the number of replicates. Under the noisiest setting, $r = 0.2$ and $\nu = 0.5$, the lowest attained size (0.07) occurs when $n = m = 300$. This is an only moderate improvement over the size (0.15) when $n = m = 50$, and is still above the nominal level. If instead the number of replicates were

Table 1. Sizes of KD and QI (in parenthesis) under 12 combinations of range, r, and smoothness, ν, and 16 combinations of sample sizes, n and m, for X and Y, respectively.

		$\nu = 0.5$				$\nu = 1.0$				$\nu = 1.5$			
n	m	$r = 0.2$	0.3	0.4	0.5	0.2	0.3	0.4	0.5	0.2	0.3	0.4	0.5
50	50	0.15	0.09	0.08	0.06	0.07	0.06	0.04	0.05	0.06	0.05	0.05	0.06
		(0.24)	(0.17)	(0.15)	(0.13)	(0.13)	(0.13)	(0.11)	(0.09)	(0.13)	(0.10)	(0.09)	(0.08)
50	100	0.13	0.10	0.07	0.06	0.07	0.06	0.04	0.05	0.06	0.06	0.05	0.06
		(0.28)	(0.21)	(0.17)	(0.13)	(0.17)	(0.14)	(0.11)	(0.10)	(0.14)	(0.11)	(0.10)	(0.09)
50	200	0.16	0.10	0.07	0.07	0.08	0.06	0.06	0.05	0.07	0.05	0.06	0.05
		(0.32)	(0.26)	(0.19)	(0.16)	(0.20)	(0.13)	(0.13)	(0.10)	(0.16)	(0.11)	(0.11)	(0.09)
50	300	0.14	0.09	0.08	0.06	0.07	0.06	0.06	0.05	0.06	0.06	0.05	0.04
		(0.39)	(0.23)	(0.19)	(0.16)	(0.19)	(0.13)	(0.13)	(0.11)	(0.16)	(0.12)	(0.11)	(0.09)
100	50	0.15	0.10	0.07	0.06	0.09	0.07	0.06	0.05	0.06	0.05	0.05	0.04
		(0.15)	(0.11)	(0.09)	(0.08)	(0.10)	(0.08)	(0.07)	(0.07)	(0.09)	(0.07)	(0.07)	(0.06)
100	100	0.10	0.07	0.07	0.05	0.06	0.05	0.04	0.05	0.06	0.04	0.04	0.04
		(0.18)	(0.13)	(0.12)	(0.11)	(0.11)	(0.09)	(0.08)	(0.08)	(0.11)	(0.08)	(0.08)	(0.07)
100	200	0.11	0.07	0.07	0.06	0.06	0.05	0.06	0.05	0.06	0.05	0.05	0.05
		(0.22)	(0.15)	(0.11)	(0.12)	(0.12)	(0.10)	(0.10)	(0.10)	(0.10)	(0.08)	(0.08)	(0.08)
100	300	0.11	0.06	0.06	0.06	0.07	0.06	0.05	0.05	0.06	0.05	0.05	0.04
		(0.22)	(0.16)	(0.14)	(0.10)	(0.15)	(0.11)	(0.08)	(0.08)	(0.11)	(0.09)	(0.09)	(0.07)
200	50	0.15	0.10	0.07	0.08	0.08	0.06	0.06	0.06	0.06	0.06	0.05	0.05
		(0.09)	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)	(0.06)	(0.06)	(0.07)	(0.07)	(0.06)	(0.06)
200	100	0.10	0.08	0.06	0.06	0.07	0.06	0.05	0.05	0.06	0.04	0.05	0.04
		(0.11)	(0.09)	(0.07)	(0.06)	(0.09)	(0.07)	(0.07)	(0.06)	(0.08)	(0.06)	(0.07)	(0.06)
200	200	0.08	0.06	0.05	0.05	0.06	0.05	0.04	0.05	0.05	0.05	0.06	0.04
		(0.13)	(0.10)	(0.08)	(0.08)	(0.09)	(0.08)	(0.08)	(0.07)	(0.08)	(0.08)	(0.07)	(0.07)
200	300	0.08	0.07	0.06	0.06	0.06	0.06	0.05	0.05	0.06	0.05	0.06	0.06
		(0.12)	(0.11)	(0.10)	(0.09)	(0.09)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
300	50	0.14	0.08	0.06	0.06	0.07	0.06	0.06	0.05	0.06	0.06	0.05	0.05
		(0.07)	(0.06)	(0.06)	(0.05)	(0.06)	(0.07)	(0.06)	(0.06)	(0.06)	(0.07)	(0.06)	(0.06)
300	100	0.10	0.06	0.05	0.05	0.06	0.06	0.04	0.05	0.05	0.05	0.04	0.05
		(0.09)	(0.08)	(0.07)	(0.06)	(0.06)	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)	(0.05)
300	200	0.09	0.06	0.06	0.07	0.07	0.05	0.04	0.04	0.06	0.05	0.05	0.05
		(0.10)	(0.07)	(0.07)	(0.08)	(0.07)	(0.07)	(0.06)	(0.06)	(0.08)	(0.06)	(0.07)	(0.07)
300	300	0.07	0.06	0.06	0.06	0.05	0.05	0.05	0.04	0.05	0.05	0.06	0.05
		(0.11)	(0.09)	(0.09)	(0.08)	(0.08)	(0.08)	(0.07)	(0.06)	(0.06)	(0.07)	(0.06)	(0.07)

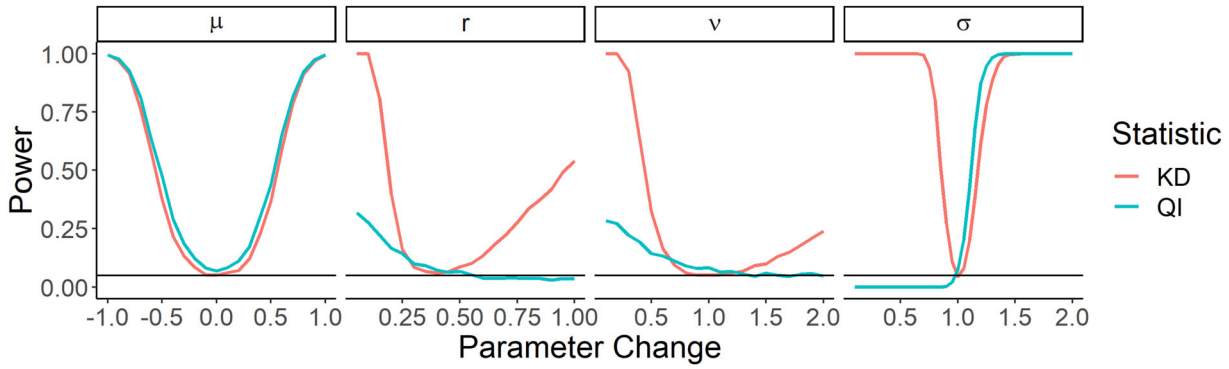


Figure 5. Power of KD and QI in detecting changes in the four parameters in the Gaussian process. Mean, range, and smoothness are presented as shifts of parameters in Y from X. Standard deviation is presented as a multiple of standard deviation in X.

fixed at $n = m = 50$ but the range and smoothness increased to either 0.5 and 1.0, respectively, then the size is controlled at the nominal level and thereafter. As with the convergence simulations though, these minimal smoothness conditions are not all that impactful in practice because functions are typically smoothed before analysis. Moreover, the sizes are stable at the nominal level once the range exceeds a threshold between 0.3 and 0.4 or $\nu \geq 1.0$ for the spatial domain $[0, 1] \times [0, 1]$. The QI test appears to inflate the size in nearly all cases compared to our test.

We compare the power of our test and the QI test in detecting changes in the four parameters μ, σ, r, and ν that govern the underlying Gaussian process in our data generation. We set the number of replicates to $n = 100$ and $m = 50$ for the two samples X and Y, respectively, and sample the functional observations in X from a Gaussian process with $r = 0.4$, $\nu = 1$, $\mu = 0$, and $\sigma = 1$. This setup ensures that the sizes of KD and QI are similar (see Table 1) so that their power functions are comparable. To generate samples of Y we let each of the parameters in Y vary around the parameter values in X. The mean, μ, was set from -1 to 1 in 0.1 increments, σ was set between 0.1 and 2 in 0.05 increments, r from 0.05 to 1 in 0.05 increments, and ν from 0.1 to 2 in 0.1 increments. This gave a total of 96 alternative models because the parameters were varied individually. The power of KD and QI were then calculated under each of these alternative models using 2000 simulations each.

Figure 5 shows the empirical power functions for both KD and QI on each parameter. Both tests are almost equally

powerful in detecting mean changes and increases in standard deviation. However, KD shows a strong improvement over QI in detecting changes in range, smoothness, and decreases in standard deviation. Two caveats about the power functions should be noted. The first is that there is a slight advantage to QI in the testing of mean, range and smoothness because QI still appears to slightly inflate size, which can be seen by observing its power function at the null value of these three parameters. The second caveat is that QI was not designed to detect decreases in standard deviation because the application for which it was designed found a drop in standard deviation desirable.

Further simulation results comparing against the FAD and the BAND are available in the supplementary materials. While FAD shows considerable power in detecting mean changes it falls short of the other methods for detecting variance changes. On average, our method maintained the highest power across the different parameters.

The situation where the mean or variance is shifted uniformly over the entire domain of the function may be a little too simplified. A more realistic scenario is that the mean, variance, and other aspects of the distribution differ heterogeneously; higher in some regions and lower in others. To study this situation, we conduct another set of simulations where the mean and variance are both allowed to vary non-uniformly over the domain, though the range and smoothness are kept constant throughout at $r = 0.4$ and $\nu = 1.0$. More specifically, we generate the mean and standard deviation of Y as two dimensional sine waves centered about 0 and 1, respectively. Then, we slowly increase the amplitude of sine waves to make X and Y deviate more in their parameters. The two sine waves are as follows:

$$\mu(s) = \left(\frac{\kappa}{2}\sin(4\pi s_1 - \pi/2) + 1\right) \otimes \left(\frac{\kappa}{2}\sin(4\pi s_1 - \pi/2) + 1\right) - 1,$$
$$\sigma(s) = \left(\frac{\kappa}{2}\sin(4\pi s_1 - \pi/2) + 1\right) \otimes \left(\frac{\kappa}{2}\sin(4\pi s_1 - \pi/2) + 1\right),$$

where $s = (s_1, s_2) \in [0,1] \times [0,1]$, κ is set to vary from 0.05 to 1 in increments of 0.05, and $\otimes$ is the Kronecker product. We fixed the number of replicates to $n = 100$ and $m = 50$ and again used 2000 simulations per κ value to estimate the power at κ.

Figure 6 shows the power functions of KD and QI under heterogeneous mean and standard deviation changes. For detecting mean changes, both KD and QI maintain comparable powers although our test carries more power than the QI test at certain ranges of mean change. It is worth noting that the power curves in this setting appear to be similar to those under the homogeneous mean change which indicates no serious power loss when the mean change is heterogeneous. A huge difference between KD and QI is observed, however, when the standard deviation change is heterogeneous: KD still maintains its power while QI seems to lose power.

Further simulations are again made available in the supplementary materials comparing KD, QI, FAD, and BAND in the heterogeneous case. FAD is extremely powerful in detecting mean changes but like QI and BAND it loses all power in detecting heterogeneous variance changes. Our method again shows the highest average power of the four approaches.

4. Evaluating Proxy Influence in Assimilated CFRs

We now apply the proposed KD statistic to evaluate the influence of proxies on the 2 m surface temperature reconstruction by examining the differences between the background and analysis states in PHYDA. In our experiment, the background state consists of a single 100 member ensemble of 2 m surface temperature fields that are randomly sampled from a single climate model simulation run. For every year of the reconstruction, the analysis state consists of a 100 member ensemble of 2 m surface temperatures (the same 100 randomly drawn ensemble members are selected for both the background and analysis). We will use our KD statistic to test for distributional differences between the background ensemble and each year's analysis ensemble so as to test for proxy influence during each reconstruction year. Formally, this refers to testing the hypothesis (1) that was formulated in Section 2.1. We will then further subdivide the background and analysis states into 12 regions, corresponding with the five oceans and seven continents, and repeat our analysis on the regions separately. Finally, we investigate how correlation between regions may impact the influence of proxies at the regional level.

4.1. Global Reconstructions

Figure 7 shows the values of KD over time. A larger value of KD corresponds to a smaller p-value and more separation between the background and analysis states in their distribution. The p-values were adjusted by the Benjamini–Yekutieli procedure

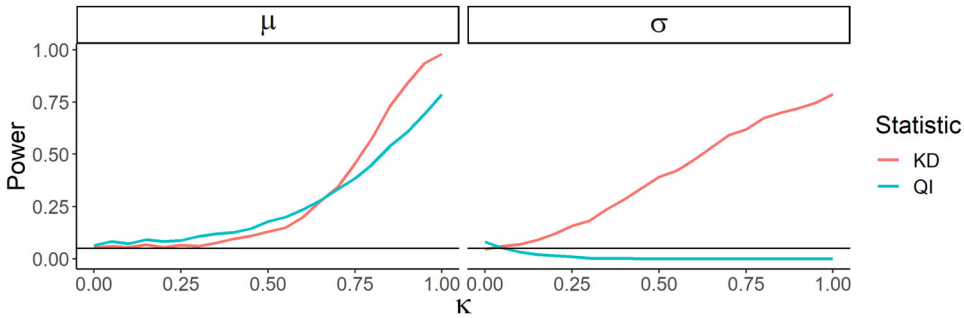


Figure 6. Power functions for KD and QI under heterogeneous differences in the mean and standard deviation between X and Y. The parameter κ controls the amplitude of the sine waves $\mu(s)$ and $\sigma(s)$, which are centered about 0 and 1, respectively.

(Benjamini and Yekutieli 2001) to have a false discovery rate (FDR) of 0.05. The uniformly near zero p-values strongly indicate that the background and analysis are significantly different in distribution each year, which suggests that the proxies indeed change the distribution of the background and have a material influence over the PHYDA reconstructions. Despite the uniformly small p-values, the magnitudes of KD indicate a relatively weak separation between the background and analysis in the beginning followed by a steadily increasing separation over time until the end of the reconstruction period. The apparent rise in separation is caused by the fact that proxy information is sequentially introduced into the reconstruction over time. Over the interval of our analysis (850 CE to 1850 CE), more proxies become available for assimilation as the reconstruction approaches 1850 CE.

4.2. Regional Variation of Proxy Influence

Analysis of the global reconstruction is important for establishing the strength of proxy influence at the global level and for confirming the upward trend of the proxy influence. A natural next step is to investigate how these effects propagate down to a regional level, namely how proxies impact the temperature reconstruction at the continental and oceanic level. Proxies are not collected uniformly across all regions as shown in Figure 8, so a weaker influence might be expected of the proxies in the poorly sampled regions than those with dense sampling. We therefore use our method to investigate the local influence of proxies.

We divide the globe into 12 regions corresponding to the five oceans and seven continents as in Figure 8. Within each of the 12 regions, we apply our test to evaluate the difference between the background and analysis states over the full reconstruction period. Analogous to the global study in Section 4.1, the progression of KD over time for all regions is summarized in Figure 9 and Figure 10. It is surprising to see that KD values over all regions share a consistent increasing trend, even for the regions with scarce proxies. Intuitively, we expect that the increasing trend holds only for the regions with abundant proxies because gradually introduced proxy information in those regions will make the analysis states more and more distinct from the background. However, due to the complex dependency structure of the climate system and its teleconnections, these regional deficits are likely being mitigated. This result intrigues us to study whether the long range dependence in the background climate states helps to stabilize the reconstruction in data-sparse regions (Figure 10).

We investigate this conjecture using the correlation maps in Figure 11, for which the value at each location describes the strongest correlation, that is, the maximum r^2, between the temperature time series at this location and every temporally available proxy location during the representative years of 1000, 1400, and 1800 CE, respectively. These maps indicate the maximum potential strength of spatial diffusion of proxy information over time. As more proxies are added toward the present, more global area becomes highly correlated with the proxy locations.

The maps in Figure 11 are helpful in understanding why some regions such as the Pacific have few proxies but show a high degree of divergence between their background and analysis states. The Pacific is strongly correlated with nearby continental regions such as North America, which has many

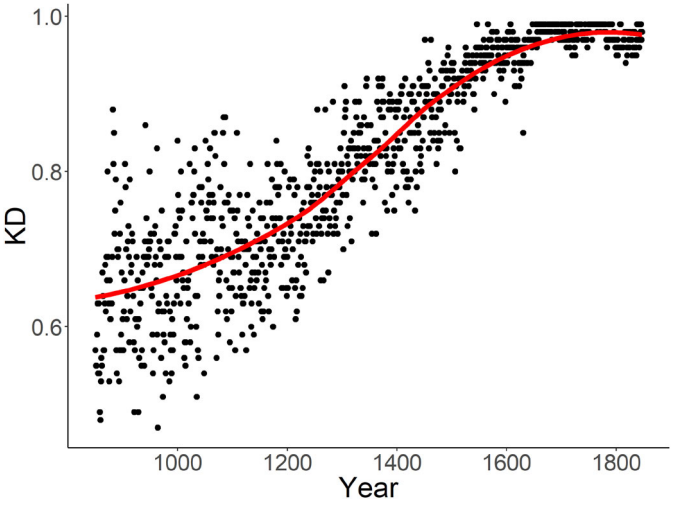
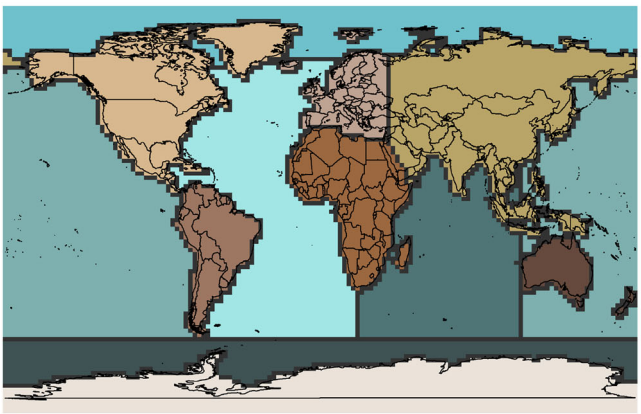


Figure 7. Value of KD over the reconstruction period from 850 CE to 1850 CE. Larger values of KD indicate larger differences between the distribution of the background and analysis states. Red line shows the overall increasing trend of KD. All p-values were less than 6×10^{-10} (after FDR adjustment).



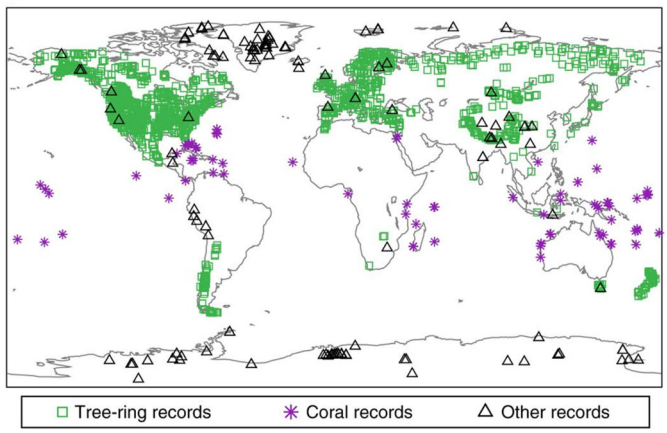


Figure 8. The left panel divides the whole globe into 12 regions marked by different colors and the right panel shows the locations of all of the $n = 2978$ proxies used in PHYDA, as in Figure 1a of Steiger et al. (2018). The vast majority of proxies are collected in North America and Europe. Not all displayed proxies are available every year in the reconstruction. More proxies become available as the reconstruction approaches the present day, see Figure 1.

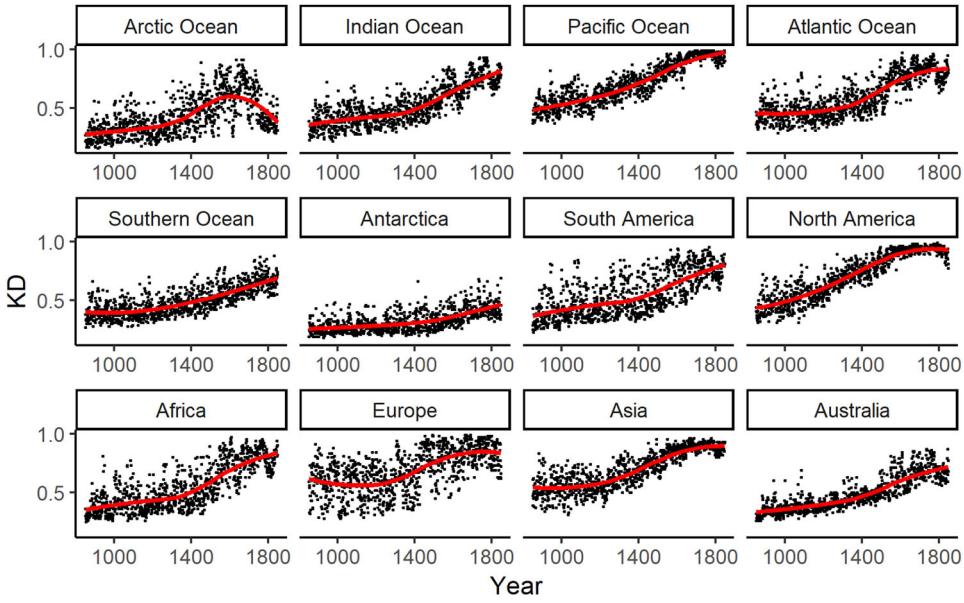


Figure 9. KD over time by region. Regional KD values were computed by measuring differences in the regions of interest within the global reconstructions. They generally follow the pattern of the global KD values with the exception of the Arctic Ocean. Red lines show the regional trends of KD.

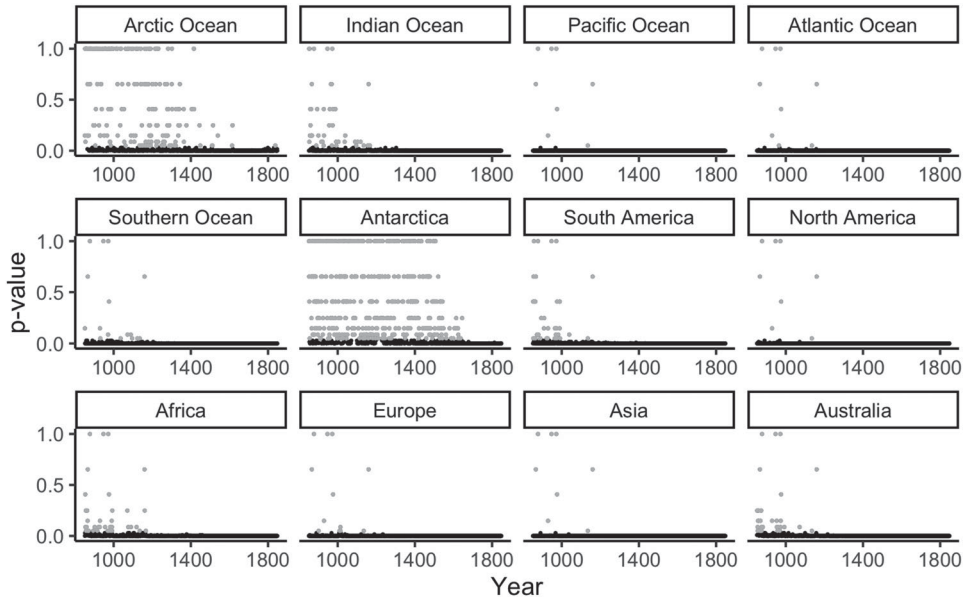


Figure 10. *p*-values of KD over time by region. Gray points indicate *p*-values over 0.05 after the Benjamini–Yekutieli FDR adjustment. Except for in the early years, most regions have statistically significant differences between their background and analysis ensembles across the reconstruction. The Arctic Ocean and Antarctica fail to reject in many more cases due to their relatively small size and lack of proxies.

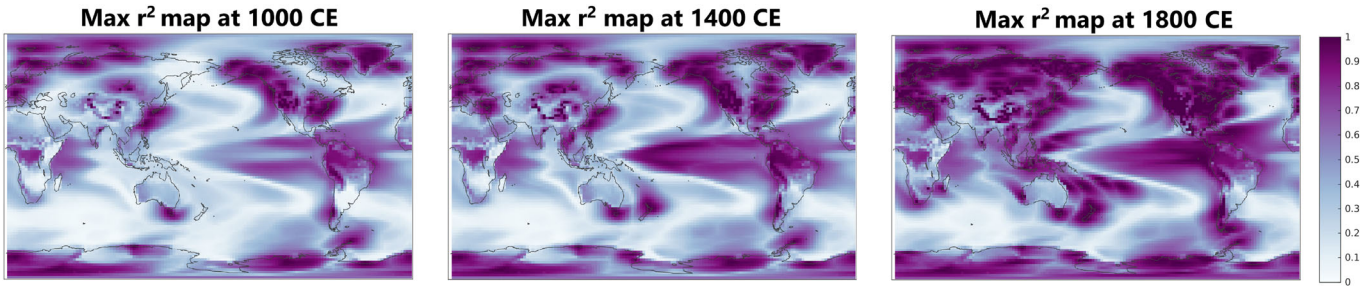


Figure 11. Proxy-point r^2 maps for representative years 1000, 1400, and 1800 CE. There is an overall increasing proxy point correlation (purple) in time. This is reflected in the increasing effect sizes seen in regions with little proxy representation.

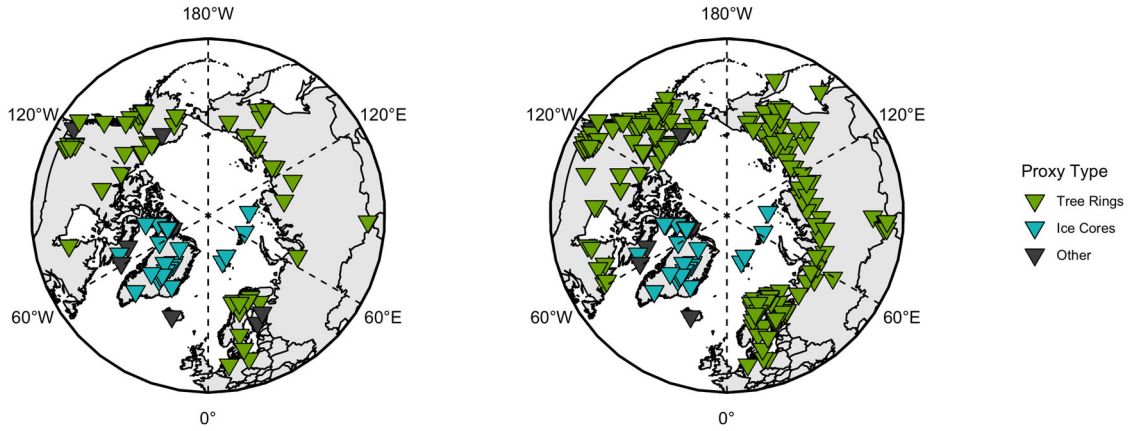


Figure 12. Left: Polar plot of proxy sites near the Arctic Ocean (above 50° latitude) in 1500 CE. Right: Proxy sites near the Arctic Ocean in 1700 CE. The number of tree-rings (green) more than triples from 1500 to 1700 CE (75 to 260 sites, respectively) without a similar increase in the number of ice cores, marine cores, corals, etc.

proxies, due to the El Niño-Southern Oscillation phenomenon. Conversely, Australia, which has few proxies and weak proxy correlations, shows a correspondingly low degree of divergence between its background and analysis states. Regions with densely sampled proxies tend to show high proxy correlation due to proximity with their own proxies, and also a high degree of background-analysis separation, as expected.

The Arctic and Southern oceans represent two anomalies with regards to their apparent proxy information. Most notably around 1600 CE the Arctic ocean experiences a strong trend reversal in KD, just when other regions are experiencing trend increases. This runs counter to the fact that both the number of proxies and the proxy-point correlations shown in Figure 11 are increasing in the Arctic over this time period.

One possible explanation of this effect is the dramatic increase in the number of Arctic tree-ring records beginning around 1600 CE, prior to which ice core and sediment records dominate in the Arctic region (see Figure 12). These latter records are isotope based and have been shown to sample far-field temperature signals across regions of the Arctic Ocean (Steiger et al. 2017), while several of the isotope records included in the reconstruction are specifically marine based. It is possible that the isotope records are therefore better samples of the far-field and exclusively marine temperatures that are reconstructed across the Arctic Ocean, relative to the land-based tree-ring records that begin to dominate in the 17th century and sample local temperature conditions. The inclusion of more widely abundant tree-ring records that are potentially less informative of far-field marine temperatures over the Arctic Ocean may therefore effectively increase the noise in the reconstruction over that region, thus obscuring the contribution from the isotope records and increasing the reliance of the DA on the information from the prior.

Conversely, the Southern ocean has relatively large values of KD when the correlation information in Figure 11 would lead us to believe that they should be much smaller. The Southern ocean has no local proxies and perhaps has the weakest overall proxy-point correlation strength, yet it experiences a strong and significant divergence between its background and analysis states. This unexpectedly strong divergence may be due to the large amount of moderate correlations observed near the Pacific, along the coast of Antarctica, and off the southern tip of South America. The cumulative effect of those moderate correlations may lead to the results in Figure 9.

5. Discussion and Conclusions

Motivated by the newly available PHYDA reconstruction product, we developed a nonparametric statistical test to compare the distributions of the ensembles in the background states and in the analysis states. The PHYDA data product was derived using a DA scheme that merges information from climate model simulations and climate proxies, the latter of which is expected to provide its due influence on the derived analysis fields. However, the nature of the DA approach and the variation of proxy information through time makes it difficult to assess the degree to which the proxies influence the final analysis product.

Optimally adding proxy information is one of the principal qualities of a DA-based reconstruction and thus knowing the cumulative effect of adding proxies is of fundamental importance, particularly as DA becomes increasingly popular (Franke, Werner, and Donner 2017; Steiger et al. 2018; Tardif et al. 2019). Before now, testing for significant proxy influence over DA reconstructions has been conducted empirically through validation procedures (e.g., Hakim et al. 2016; Singh et al. 2018). Our test instead provides a direct and powerful way to formally quantify the information added by proxies to the analysis states based on changes in their distribution from the background. By treating each ensemble member in the background and analysis states as continuous two dimensional surfaces, our test statistic based on functional data depth is able to measure the difference in distribution between ensembles in the two states.

Due to the nonparametric nature of functional data depth, our method does not require any distributional or model assumptions on the observations. Our method also does not require that the curves be square integrable, second-order stationary, or even strictly continuous. We showed numerically that the asymptotic distribution of our test statistic converges to or at least is well approximated by the Kolmogorov distribution. Additional simulations in the supplementary materials show the same conclusion even when the spatial data are from a non-Gaussian process (see Figures 3 and 4 in the supplementary materials). We also demonstrated that the sizes of our test are

well controlled near the nominal level, even under moderate sample sizes, and that our test's powers are highly competitive with the QI test in Liu and Singh (1993).

Our results provide strong evidence of a clear divergence between the background and analysis states associated with PHYDA. The degree of separation, however, depends greatly on geographical location and time period. An overall upward trend in proxy influence is seen and it is generally maintained even when subdividing the globe into oceanic and continental sub-regions. With the notable exception of the Arctic, these findings are consistent with the fact that proxy information steadily increases as the reconstruction period approaches the present day. This confirms that increasing proxy information is associated with commensurate influence on the assimilated reconstructions, which suggests that the influence of the model prior is minimized as proxy networks become considerably more dense, therefore placing less emphasis on which model should be used to form the prior. This also suggests that more proxies should be collected further back in time to improve reconstruction skill over all parts of the Common Era.

We have also found that, despite the stark imbalance in proxy density in the different geographic regions, most regions exhibit an increasing separation between the background and analysis states. The mitigating effect for the proxy deficit regions is mostly attributable to the long-range dependency structure that proxies and temperatures often display. Some regions such as the Pacific Ocean and South America have very few local proxies but due to their strong overall correlation with other regions they still benefit from proxies collected remotely. These results therefore suggest that the desirable addition of proxy information to data assimilated reconstructions extends beyond the immediate regions where proxies are densely sampled. This provides credence to the idea that the geographic regions outside of dense proxy sampling may still establish some reconstruction skill, particularly in the last several centuries before the present.

In addition to the important results for assessing assimilated reconstruction products, our test is much more broadly applicable. Our generic formulation allows it to be applied to any functional data that the depth function can handle, including curves on $\mathbb{R}$ and higher dimensional functions on $\mathbb{R}^n$. In our framework, each X_i and Y_j can also be multivariate valued so long as they both map to the same subspace of $\mathbb{R}^p$. This only changes integration to be over a multivariate depth instead of a univariate depth. Our method can also be useful for comparing image data in medical studies, and meet the increasing demand of comparing simulated climate from different climate models and comparing the simulated climate to observations.

Data Availability

PHYDA is publicly available at the Zenodo data repository as NetCDF4 files: *https://doi.org/10.5281/zenodo.1154913*. The 100 member ensembles of PHYDA used herein are available at: *http://clifford.ldeo.columbia.edu/nsteiger/recon_output/phyda_ens/*.

Appendix

Proof of Proposition 2.1. Let P be a distribution on $C[0,1]^p$ and suppose $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_n\}$ are two iid samples from P. Let $\widehat{F}_n(\cdot)$ and $\widehat{G}_m(\cdot)$ be defined as before with each converging in distribution to F, the distribution over $D(\cdot, P)$. Let $x \in X$, then

$$\begin{aligned}
&\sqrt{\frac{nm}{n+m}} \max_{x\in X} |\widehat{F}_n(x) - \widehat{G}_m(x)| \\
&\leq \sqrt{\frac{nm}{n+m}} \max_{x\in X} |\widehat{F}_n(x) - F(x)| + \sqrt{\frac{nm}{n+m}} \max_{x\in X} |F(x) - \widehat{G}_m(x)| \\
&\simeq \sqrt{m} \max_{x\in X} |\widehat{F}_n(x) - F(x)| + \sqrt{m} \max_{x\in X} |F(x) - \widehat{G}_m(x)|.
\end{aligned}$$

Because $n \gg m$. By the enforced uniformity of $\widehat{F}_n(x)$ we get that $\max_{x\in X} |\widehat{F}_n(x) - F(x)| = o_p(\frac{1}{\sqrt{n}})$ and so the following upper bound

$$\leq o_p(1) + \sqrt{m} \max_{x\in X} |F(x) - \widehat{G}_m(x)|.$$

The second term is simply a one sample Kolmogorov–Smirnov statistic so the whole quantity converges to the Kolmogorov distribution. □

Supplementary Materials

Supplementary results: PDF file containing additional power simulations and comparisons between FAD and KD on the PHYDA data. (pdf file).

Source code: Zip file containing code to produce all figures and results in the manuscript and supplementary file. (zip file)

R-package for the KD statistic: R-package "kstat" containing code to compute the the Kolmogorov Depth statistic and p-values using the asymptotic and permutation distributions. Available through GitHub: `trevor-harris/kstat`. (GNU zipped tar file)

Funding

This research was supported in part by the National Science Foundation through awards OISE-1743738, AGS-1602581, AGS-1602920, AGS-1805490, NSF-AGS-1602845, and NSF-DMS-1830312. Dr. Narisetty was also partially supported by NSF Award DMS-1811768. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. Lamont contribution no. 8428.

References

Benjamini, Y., and Yekutieli, D. (2001), "The Control of the False Discovery Rate in Multiple Testing Under Dependency," *The Annals of Statistics*, 29, 1165–1188. [10]

Benko, M., Hardle, W., and Kneip, A. (2009), "Common Functional Principal Components," *The Annals of Statistics*, 37, 1–34. [2]

Briggs, W., and Levine, R. (1997), "Wavelets and Field Forecast Verification," *Monthly Weather Review*, 125, 1329–1341. [2]

Christiansen, B., and Ljungqvist, F. C. (2017), "Challenges and Perspectives for Large-Scale Temperature Reconstructions of the Past Two Millennia," *Reviews of Geophysics*, 55, 40–96. [1]

Corain, L., Melas, V., Pepelyshev, A., and Salmaso, L. (2014), "New Insights on Permutation Approach for Hypothesis Testing on Functional Data," *Advances in Data Analysis and Classification*, 8, 339–356. [2]

Cuevas, A., and Fraiman, R. (2009), "On Depth Measures and Dual Statistics. A Methodology for Dealing With General Data," *Journal of Multivariate Analysis*, 100, 753–766. [3,4]

Fraiman, R., and Muniz, G. (2001), "Trimmed Means for Functional Data," *Test*, 10, 419–440. [4]

Franke, J. G., Werner, J. P., and Donner, R. V. (2017), "Reconstructing Late Holocene North Atlantic Atmospheric Circulation Changes Using Functional Paleoclimate Networks," *Climate of the Past*, 13, 1593–1608. [12]

Good, P. (2013), *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, New York: Springer-Verlag. [6]

Goosse, H., Crespin, E., Dubinkina, S., Loutre, M.-F., Mann, M. E., Renssen, H., Sallaz-Damaz, Y., and Shindell, D. (2012), "The Role of Forcing and Internal Dynamics in Explaining the Medieval Climate Anomaly," *Climate Dynamics*, 39, 2847–2866. [1]

Guillot, D., Rajaratnam, B., and Emile-Geay, J. (2015), "Statistical Paleoclimate Reconstructions via Markov Random Fields," *The Annals of Applied Statistics*, 9, 324–352. [1]

Hakim, G., Emile-Geay, J., Steig, E., Noone, D., Anderson, D., Tardif, R., Steiger, N., and Perkins, W. (2016), "The Last Millennium Climate Reanalysis Project: Framework and First Results," *Journal of Geophysical Research: Atmospheres*, 121, 6745–6764. [1,2,12]

Hall, P., and Van Keilegom, I. (2007), "Two Sample Tests in Functional Data Analysis Starting From Discrete Data," *Statistica Sinica*, 17, 1511–1531. [2]

Hering, A., and Genton, M. (2011), "Comparing Spatial Predictions," *Technometrics*, 53, 414–425. [2]

Horvath, L., Kokoszka, P., and Reeder, R. (2013), "Estimation of the Mean of Functional Time Series and a Two Sample Problem," *Journal of the Royal Statistical Society*, Series B, 75, 103–122. [2]

Jones, P., Briffa, K., Osborn, T., Lough, J., van Ommen, T., Vinther, B., Luterbacher, J., Wahl, E., Zwiers, F., Mann, M., Schmidt, G., Ammann, C., Buckley, B., Cobb, K., Esper, J., Goosse, H., Graham, N., Jansen, E., Kiefer, T., Kull, C., Kttel, M., Mosley-Thompson, E., Overpeck, J., Riedwyl, N., Schulz, M., Tudhope, A., Villalba, R., Wanner, H., Wolff, E., and Xoplaki, E. (2009), "High-Resolution Palaeoclimatology of the Last Millennium: A Review of Current Status and Future Prospects," *The Holocene*, 19, 3–49. [1]

Lee, T. C., Zwiers, F. W., and Tsao, M. (2008), "Evaluation of Proxy-Based Millennial Reconstruction Methods," *Climate Dynamics*, 31, 263–281. [1]

Li, B., and Smerdon, J. (2012), "Defining Spatial Comparison Metrics for Evaluation of Paleoclimatic Field Reconstructions of the Common Era," *Environmetrics*, 23, 394–406. [2]

Li, B., Zhang, X., and Smerdon, J. (2016), "Comparison Between Spatiotemporal Random Processes and Application to Climate Model Data," *Environmetrics*, 27, 267–279. [1,2]

Liu, R. (1990), "On a Notion of Data Depth Based on Random Simplices," *The Annals of Statistics*, 18, 405–414. [2]

Liu, R., and Singh, K. (1993), "A Quality Index Based on Data Depth and Multivariate Rank Test," *Journal of the American Statistical Association*, 88, 252–260. [2,3,5,6,13]

Lopez-Pintado, S., and Romo, J. (2009), "On the Concept of Depth for Functional Data," *Journal of the American Statistical Association*, 104, 718–734. [2,4]

Lund, R., and Li, B. (2009), "Revisiting Climate Region Definitions via Clustering," *Journal of Climate*, 22, 1787–1800. [2]

Mann, M. E., Bradley, R. S., and Hughes, M. K. (1998), "Global-Scale Temperature Patterns and Climate Forcing Over the Past Six Centuries," *Nature*, 392, 779–787. [1]

Masson-Delmotte, V., Schulz, M., Abe-Ouchi, A., Beer, J., Ganopolski, J., González Rouco, J. F., Jansen, E., Lambeck, K., Luterbacher, J., Naish, T., Osborn, T., Otto-Bliesner, B., Quinn, T., Ramesh, R., Rojas, M., Shao, X., and Timmermann, A. (2013), *Information From Paleoclimate Archives*, Cambridge, UK: Cambridge University Press, pp. 383–464. [1]

Mosler, K., and Polyakova, Y. (2012), "General Notions of Depth for Functional Data," arXiv no. 1208.1981. [4]

Nagy, S., Gijbels, I., Omelka, M., and Hlubinka, D. (2016), "Integrated Depth for Functional Data: Statistical Properties and Consistency," *European Series in Applied and Industrial Mathematics: Probability and Statistics*, 20, 95–130. [3,4]

Narisetty, N., and Nair, V. (2017), "Extremal Depth for Functional Data and Applications," *Journal of the American Statistical Association*, 111, 1705–1714. [4]

Otto-Bliesner, B. L., Brady, E. C., Fasullo, J., Jahn, A., Landrum, L., Stevenson, S., Rosenbloom, N., Mai, A., and Strand, G. (2016), "Climate Variability and Change Since 850 CE: An Ensemble Approach With the Community Earth System Model," *Bulletin of the American Meteorological Society*, 97, 735–754. [3]

Pavlicova, M., Santer, T., and Cressie, N. (2008), "Detecting Signals in FMRI Data Using Powerful FDR Procedures," *Statistics and Its Interface*, 1, 23–32. [2]

Pomann, G., Staicu, A., and Ghosh, S. (2016), "A Two Sample Distribution Free Test for Functional Data With Application to a Diffusion Tensor Imaging Study of Multiple Sclerosis," *Journal of the Royal Statistical Society*, Series C, 65, 395–414. [2]

Ramsay, J., and Silverman, B. (2005), *Functional Data Analysis*, New York: Springer-Verlag. [2]

Shen, X., Huang, H., and Cressie, N. (2002), "Nonparametric Hypothesis Testing for a Spatial Signal," *Journal of the American Statistical Association*, 97, 1122–1140. [2]

Singh, H. K. A., Hakim, G. J., Tardif, R., Emile-Geay, J., and Noone, D. C. (2018), "Insights Into Atlantic Multidecadal Variability Using the Last Millennium Reanalysis Framework," *Climate of the Past*, 14, 157–174. [2,12]

Smerdon, J. E. (2012), "Climate Models as a Test Bed for Climate Reconstruction Methods: Pseudoproxy Experiments," *Wiley Interdisciplinary Reviews: Climate Change*, 3, 63–77. [1]

Smerdon, J. E., and Pollack, H. N. (2016), "Reconstructing Earth's Surface Temperature Over the Past 2000 Years: The Science Behind the Headlines," *Wiley Interdisciplinary Reviews: Climate Change*, 7, 746–771. [1]

Snell, S., Gopal, S., and Kaufmann, R. (2000), "Spatial Interpolation of Surface Air Temperatures Using Artificial Neural Networks: Evaluating Their Use for Downscaling GCMs," *Journal of Climate*, 13, 886–895. [2]

Staicu, A., Li, Y., Crainiceanu, C., and Ruppert, D. (2014), "Likelihood Ratio Tests for Dependent Data With Applications to Longitudinal and Functional Data Analysis," *Scandinavian Journal of Statistics*, 41, 932–949. [2]

Steiger, N. J., Hakim, G. J., Steig, E. J., Battisti, D. S., and Roe, G. H. (2014), "Assimilation of Time-Averaged Pseudoproxies for Climate Reconstruction," *Journal of Climate*, 27, 426–441. [1,2]

Steiger, N. J., Smerdon, J. E., Cook, E. R., and Cook, B. I. (2018), "A Reconstruction of Global Hydroclimate and Dynamical Variables Over the Common Era," *Scientific Data*, 5, 180086. [1,2,3,10,12]

Steiger, N. J., Steig, E. J., Dee, S. G., Roe, G. H., and Hakim, G. J. (2017), "Climate Reconstruction Using Data Assimilation of Water Isotope Ratios From Ice Cores," *Journal of Geophysical Research: Atmospheres*, 122, 1545–1568. [12]

Stein, M. L. (2012), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer-Verlag. [6]

Tardif, R., Hakim, G. J., Perkins, W. A., Horlick, K. A., Erb, M. P., Emile-Geay, J., Anderson, D. M., Steig, E. J., and Noone, D. (2019), "Last Millennium Reanalysis With an Expanded Proxy Database and Seasonal Proxy Modeling," *Climate of the Past*, 15, 1251–1273. [1,12]

Tingley, M. P., Craigmile, P. F., Haran, M., Li, B., Mannshardt, E., and Rajaratnam, B. (2012), "Piecing Together the Past: Statistical Insights Into Paleoclimatic Reconstructions," *Quaternary Science Reviews*, 35, 1–22. [1]

Wang, W., Anderson, B., Entekhabi, D., Huang, D., Su, Y., Kaufmann, R., Potter, C., and Myneni, R. (2007), "Intraseasonal Interactions Between Temperature and Vegetation Over the Boreal Forests," *Earth Interactions*, 11, 1–30. [2]

Zhang, J., and Chen, J. (2007), "Statistical Inferences for Functional Data," *The Annals of Statistics*, 35, 1052–1079. [2]

Zhang, X., and Shao, X. (2015), "Two Sample Inference for the Second Order Property of Temporally Dependent Functional Data," *Bernoulli*, 21, 909–929. [2]

Zuo, Y., and He, X. (2006), "On the Limiting Distributions of Multivariate Depth-Based Rank Sum Statistics and Related Tests," *The Annals of Statistics*, 34, 2879–2896. [6]

Zuo, Y., and Serfling, R. (2000), "General Notions of Statistical Depth Function," *The Annals of Statistics*, 28, 461–482. [4]