Testing Exhangeability of Two Spatiotemporal Processes with Applications to Evaluating Proxy Influence in Data Assimilation

Trevor Harris[1], Bo Li[1], Nathan Steiger[2], Jason Smerdon[2], Naveen Narisetty[1], J. Derek Tucker[3]

May 4, 2019

Abstract

Statistical inference on spatiotemporal processes is a fundamental problem in many fields including Ecology, Oceanography, and Climatology. Of particular interest to the paleoclimate community is the study of climate field reconstructions (CFRs) with seasonal to annual resolution spanning the last several millennia. CFRs attempt to recover spatiotemporal fields of climate variables, using proxy records of past climate variability, and have emerged as important tools for studying the mechanisms of climate change. Motivated by assessing differences between CFRs, we propose a new method for evaluating the differences in the distributions of two spatiotemporal processes by using the notions of data depth and functional data. Our test is robust, computationally efficient, distribution free and and has a convenient asymptotic distribution. We apply our test to study global and regional proxy influence on a data assimilation based CFR by comparing its background and analysis states. We find that as the number of assimilated proxies increase nearer to the present, the states' distributions steadily diverge. This indicates an increasing proxy influence and that proxy influence can extend far beyond proxy sites.

Keywords: Functional depth; Exchangeability; Spatial fields

Short title: Exchangeability of Spatial Fields

[1]Department of Statistics, University of Illinois at Urbana-Champaign
[2]Lamont-Doherty Earth Observatory, Columbia University, New York City, NY
[3]Sandia National Laboratories, Albuquerque, NM

1 Introduction

Since their first high-profile application two decades ago (**?**), spatio-temporal climate field reconstructions (CFRs) have become increasingly popular in the climate science community for their ability to reconstruct global climate variability over many hundreds of years. Such CFRs are used to uncover the dominate modes of spatial and temporal variation in the atmosphere-ocean climate system. CFRs critically rely on large networks of climate proxies, such as isotopic information in ice cores and the width of tree rings, that indirectly record climate variables like temperature and moisture over time (e.g., **?**). CFRs therefore provide long-term climate information that is critical to understanding the fundamental processes of the climate system and how climate may change in the future.

A recent innovation in CRFs are Data Assimilation (DA) algorithms. Data assimilation algorithms are a class of reconstruction methods that optimally combine general circulation models (GCMs) with proxy information to create paleoclimate reconstructions (**??**). Their primary advantage over traditional reconstructions methods is their ability to jointly reconstruct multiple atmosphere-ocean variables and to do so in a physically consistent manner. Multiple, physically consistent climate variables are absolutely essential for understanding phenomena such as the Medieval megadroughts of the American West (e.g., **??**). An additional advantage of ensemble DA reconstruction algorithms is that they 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 (**?**), 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 critically important given the noisy relationship between paleoclimate proxies and climate variables.

Despite the rapid development of DA-based reconstruction methods, much is unknown about the influence of each of the their two components: climate models and paleoclimate proxies. In currently published DA-based CFRs (e.g., **?**) it is unknown how much information

the models and the proxies contribute to the end product. Furthermore, it is also unclear 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 such DA-based CFRs are fundamentally a product of the underlying climate model and not of the historical proxy data. A lack of proxy influence would therefore indicate a need to fundamentally re-evaluate DA methodologies. **[Bo: I feel it is still not very clear about what is the purpose to do this project. Is the whole purpose to evaluate DA methodology or study the influence of proxies to temperature reconstruction?]**

1.1 Reconstruction Data

[Bo: This can be absorbed into the introduction unless there is more to say about the data. In the latter case, we make it an independent section or subsection] Our DA-based reconstruction comes from Paleo Hydrodynamics Data Assimilation product (PHYDA), which is a global paleoclimate reconstruction project that reconstructs both temperature and moisture variables (**?**). PHYDA is established based on the Community Earth System Model (CESM) general circulation model which provides high resolution computer simulations of the earth's past climate. The output from CESM forms the starting point for reconstructions, i.e. the background states; then an ensemble Kalman filter is performed to optimally update the background with a network of about 3000 proxies to form the analysis states. Figure 1 shows an example of a background state and an analysis state.

It is important to note that the DA method underlying PHYDA is somewhat different from typical DA schemes (such as used in weather forecasting). Instead of propagating information forward in time by using the analysis state in year $t-1$ as the background for year t, the same exact background state **[Bo: be more specific?]** is used for each year t. **[Bo: Why perform DA in such a way?]**This implies that the differences between the analysis state and the background state in any given year are solely due to the proxy information available that year. **[Bo: Honestly I am still confused what specifically**

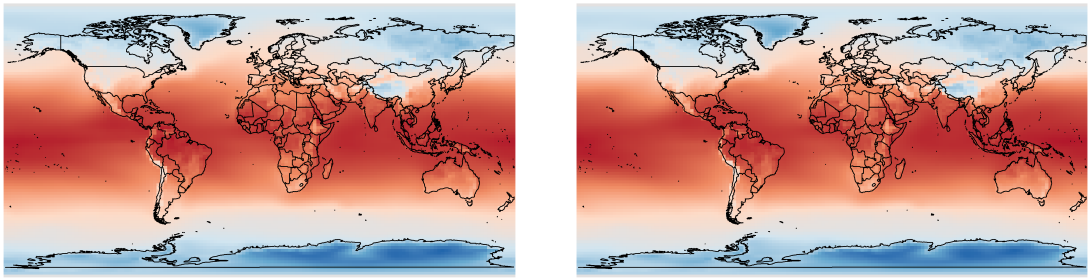

Figure 1: An example of a background field from the CESM ensemble (left) and an assimilated field in the analysis ensemble (right) in 860 CE.

have been used as background.]

We use the global, annually resolved, two meter surface temperature from PHYDA over the 1000 year interval of 850 CE to 1850 CE. Under the climatology context, annual often refers to the interval between April and March of the following year, hence we have 998 climatological years out of 1000. The fields are based on a two degree by two degree grid (144×96 grid points) spanning the entire Earth. Because of the very large data fields produced by PHYDA, we used a 100 member sub-ensemble randomly drawn from PHYDA's original 998 member ensemble reconstruction. **[Bo: You (Jason and Nathan) particularly generated this for us or that is what you did in general? Also, I think we need more details of the data. For example, what is the format of background states?]**

1.2 Previous work in random fields comparison

Comparing two spatial processes has been addressed in both geostatistics and functional data analysis literatures. The general strategy for all those methods is to reduce the dimension of the random process either by a low rank decomposition or by parameterization, and then develop the test for comparison in the reduced dimension.

From the geostatistical point of view, wavelet decomposition has been used to reduce the stochastic process to a finite number of wavelet coefficients and thus the comparison

between two processes is transformed to evaluating the difference between two sets of wavelet coefficients (**???**). **?** and **?** extended the method of comparing two time series by evaluating their average loss differentials over the forecast error to assessing the loss function of spatial interpolations. Later **?** developed a test that can handle more arbitrary and user specified loss functions. Parametric methods have been mainly focusing on testing equality of the first and second moments. Motivated by **?** that compared two time series, **?** proposed a parametric method of jointly assessing the first two moments between two random fields.

Functional data analysis approaches assume the spatial random fields to be noisy realizations of an underlying continuous function observed on a finite grid. Traditional dimension reduction methods for functional data such as functional principal component analysis are usually employed in the method development for comparing two random fields. To date, most functional approaches have focused on testing equality of the mean functions arising from two functional data sets (**????**), although lately the second order structure of functional data also received its due attention (**?**). **?** extended **?** to evaluate the joint difference in mean and covariance structure as well as in trend surface between two spatiotemporal random fields. A nice feature of functional data analysis approaches as opposed to geostatistical methods is that usually assumptions about distribution and model specification are relaxed though at the price of requiring replicates with functional data.

All the above procedures are however inadequate to our problem, because the proxies can affect both mean and covariance structure and higher order structure of the climate field **[Bo: Is DA able to alter the higher order structure of climate field?]**. Furthermore, the rich ensembles from both the background state and the analysis state allow us to examine more information than mean and covariance. Therefore, we aim to compare the distribution of two ensembles to identify the change caused by proxy. To take advantage of ensembles, we will employ a functional data approach that is both distribution and parameter free. Until recently the problem of comparing distribution of functions has remained relatively untouched. **?** proposed a Cramer-von Mises-like test by constructing an empirical distribution over each

of the samples and measuring the L^2 distance between the empirical distributions. Later **?** introduced a permutation test on the leading coefficients of the common functional principal components (FPCs) and **?** introduced three omnibus tests for combining pointwise tests on the observations of the functions. Each of these methods depends on a resampling procedure which renders them computationally prohibitive for data assimilation output.**?** proposed a method based on marginal FPCs that does not require resampling. Their test compares the distributions of the marginal FPCs using the two sample Anderson-Darling test and a Bonferroni correction. We initially tried a method similar to theirs but found that the principal components in our data had nearly degenerate distributions and thus failed for being used as a valid test. We conjecture that this could be because the impact of proxies is only limited to local scale and therefore it is difficult for the dominant eigenfunctions to register.

The test we propose is based on the functional data analysis paradigm, but is conceptually different from previous efforts. We will use the concept of data depth to construct a nonparametric statistic for assessing the equality of distributions of two functional data sets. This manner of testing has been explored by **?** who introduced the Quality Index (QI) for comparing two multivariate distributions. The Quality index essentially measured the mean outlyingness of one sample from another using data depth. Their asymptotic results were later finalized in **?** who proved that the asymptotic distribution of the two sample QI was Gaussian. We will extend their ideas to the functional setting and propose a modification that makes our statistic invariant to the reference distribution. The use of depth, and particularly integrated Tukey depth, ensures our test to be computationally efficient, distribution free, and invariant to location, scale, warping, and other nuisance properties that could influence the testing (**?**).

The rest of the paper will proceed as follows. Section 2 formulate the problem of evaluating proxy influence into a statistic question and proposes a new test statistic for assessing the exchangeability of two sets of functions. Section 3 shows simulation results to validate the asymptotic behavior of our new test statistic and evaluate the size and power of our

proposed test. Finally in Section 4 we apply it to the data assimilation model at both the regional and global level and answer the questions laid on in the introduction. Section 5 is a small discussion on the results and some directions for future work.

2 Statistical Solution

2.1 Formulation of evaluating proxy influence

Suppose X and Y_t represent the ensemble in the background state and the ensemble in the analysis state at time t in our DA reconstruction, respectively. Under our assimilation design, the proxies at time t are the only contributor to the differences between the two sets of ensembles. Our goal then is to define and quantify the differences between X and Y_t at each year in order to quantify the proxy influence.

How much impact that proxies may have on analysis states depend 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 (**?**). As shown in Fig. 1, the effects from proxies may be subtle and hard to visually detect because they may be thinly diffused over a non-contiguous area due to spatial correlations and teleconnections. It seems unlikely that the totality of their effects would be completely described by changes in the first two moments of the climate field. A more comprehensive approach would instead be to test for changes in the distributions of X and Y_t. We thus formulate our problem into the following sequence of hypotheses ([**Bo: sounds multiple testing is involved**]):

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

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 to those outcomes at t increases. Since proxies are progressively added into the background states, we may expect

that differences between the two distributions increase over time.

Under the functional data analysis regime we assume that the observed data are generated from continuous functions combined with additive noise, instead of from 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 $i.i.d$ realization of a stochastic process on a functional space.

Suppose we observe two sets of functional data, $X = \{(s, X_i(s))_{s\in D}\}_{i=1}^{n}$ and $Y = \{(s, Y_j(s)_{s\in D})\}_{j=1}^{m}$, where D is a compact subspace of $\mathbb{R}^p$. For simplicity, D is set to be $[0,1]^p$ and we assume each functional data is observed on the same set of grid points in $[0,1]^p$. Furthermore, we assume each function X_i and Y_j is a univariate continuous function on the domain $[0,1]^p$, i.e. $X_i : [0,1]^p \mapsto \mathbb{R}$ 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 $C[0,1]^p$. We therefore consider each functional data X_i (or Y_j) as being a random sample from a process in $C[0,1]^p$. For 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 in $C[0,1]^2$ and suppose each $X_i \sim P$ and $Y_j \sim Q$. As mentioned in section 2.1, we are interested in testing if the functional data in X and in Y follow the same distribution, then (1) is equivalent to the hypothesis

$$H_0 : P = Q; \quad H_A : P \neq Q. \tag{2}$$

We will use data depth to construct a two sample Kolmogorov-Smirnov type test. Other distribution free tests such as 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 Data depth and integrated Tukey depth

Data depth is a statistical concept for quantifying the "centralness" or "depth" of the observed data points with respect to a reference distribution. The closer an observation is located to the median of the distribution the more central it is and hence the higher its depth value. Since the reference distribution is typically unknown the depth has to be estimated via depth functions. Many depth functions have been developed for functional data including the integrated band depth (**?**), extremal depth (**?**), and various integrated depths (**?**). Each of these depth functions has its own strengths and weaknesses but none dominates the others in all aspects, see **?** and **?** for a review. We choose the integrated depth as the basis of our test for its simplicity, computational tractability, and highly desirable theoretical properties.

Integrated depths are a well studied class of functional data depth measures that are first introduced by **?** and then studied extensively by **?** and **?**. Integrated depths are defined in two stages. First, a univariate depth function is 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 and the simplicial depth are perhaps the two most popular ones. They are closely related and both their integrated versions come with strong theoretical guarantees. We opted to use the Tukey depth but the simplicial depth would have been equally effective since the orderings they induce are nearly identical.

The integrated Tukey depth is defined as follows. Let u be an element of $\mathbb{R}$ and let F be an absolutely continuous distribution on $\mathbb{R}$. The univariate Tukey depth of u with respect to F is

$$D_T(u, F) = \min\{F(u), (1 - F(u-))\}.$$

To enforce the depth function $D(u, F)$ within the range of $[0, 1]$, we scale $D_T(u, F)$ by 2 and define the scaled depth as

$$D(u, F) = 2D_T(u, F) = 1 - |1 - 2F(u)|$$

for a continuous F. We can further allow F, and consequently $D(\cdot, F)$, to depend on the location t. Let $X \in C[0, 1]^p$, the univariate Tukey depth of X at $t \in [0, 1]$ thus immediately follows as

$$D(X(t), F_t) = 1 - |1 - 2F_t(X(t))|,$$

where F_t is the distribution of $X(t)$. We then define the integrated Tukey depth of X, with respect to P in (2), as

$$D(X, P) = \int_0^1 D(X(t), F_t)dt.$$

To ensure that this depth function is proper we refer to the desirable criteria proposed by **?** and later by **?**. In **?** 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 of the sample depths. These properties broadly assure us that as a center-out ordering the integrated depth is well behaved.

2.3 Test statistic

Based on the data depth we propose a test statistic K for our hypothesis (2). Basically, K measures the outlyingness of either P over Q or Q over P, give that P and Q may not always appear mutually outlying from each other as depth only measures centrality. For example, if one of the distributions is nested inside the other then the nested distribution will not appear outlying.

Denote P_n as the empirical estimate of P based on the sample X and Q_m the empirical estimate of Q based on Y. We start by considering P_n fixed and measuring the outlyingness of Q_m over P_n. This proceeds with first defining the following two empirical measures for any given $X_k \in X$:

$$\widehat{F}_n(X_k) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(D(X_i, P_n) \leq D(X_k, P_n)) \tag{3}$$

$$\widehat{G}_m(X_k) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}(D(Y_j, P_n) \leq D(X_k, P_n)). \tag{4}$$

Essentially, $\widehat{F}_n(X_k)$ is a rescaling of the depth of X_k such that $\widehat{F}_n(X_k) = 1/n$ if $D(X_k, P_n)$ is the smallest, $2/n$ if the depth is the second smallest, and so until it reaches 1 when the depth becomes the largest. It acts as the empirical cumulative distribution function of the depths of X_k and thus naturally follows **[Bo: approximate?]** a uniform distribution. The second quantity $\widehat{G}_m(X_k)$ can be considered as a transformation of the depths of Y with respect to the depths of X. Under H_0 in (2), $\widehat{G}_m$ should be approximately uniform so a deviation of $\widehat{G}_m$ from the uniform distribution indicates an outlyingness of Q_m from P_n. The introduction of $\widehat{F}_n(X_k)$ and $\widehat{G}_m(X_k)$ 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, $\widehat{F}_n(X_k)$ and $\widehat{G}_m(X_k)$ for $k = 1, \ldots, n$. The latter can naturally be quantified using the Kolmogorov distance over the set X **[Bo: Is below correct? not sure X_k is placed correctly]**,

$$K_{P_n}(X, Y) = \max_{X_k \in X} |\widehat{F}_n(X_k) - \widehat{G}_m(X_k)|. \tag{5}$$

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

$$\widetilde{F}_n(Y_k) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(D(X_i, Q_m) \leq D(Y_k, Q_m))$$
$$\widetilde{G}_m(Y_k) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}(D(Y_j, Q_m) \leq D(Y_k, Q_m)).$$

These two quantities exactly mirror $\widehat{F}_n$ and $\widehat{G}_m$ except that now $\widetilde{G}_m$ becomes **[Bo: approximately?]** uniform and $\widetilde{F}_n$ the indicator for the outlyingness of P_n from Q_m. We again take Kolmogorov distance, but now over the set Y, as the measure of outlyingness,

$$K_{Q_m}(X, Y) = \max_{Y_k \in Y} |\widetilde{F}_n(Y_k) - \widetilde{G}_m(Y_k)|.$$

To define the overall test statistic K we take the maximum of the two distances:

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

The test statistic $K(X, Y)$ attains a level of symmetry by making the test invariant to the reference distribution. It is strictly non-negative and it equals 0 only under H_0 in the hypothesis (2). Thus the originally stated hypothesis (1) can be tested by evaluating how significantly $K(X, Y)$ is greater than 0. We expect $K(X, Y)$ to detect the difference between P and Q in either mean, scale, or correlation structures, for both situations where there are global shifts in the parameters as well as where parameters change randomly over the domain. The biggest difference between our test statistic $K(X, Y)$ and the Quality Index (QI) in **[Bo: reference]** is that our test does not depend on a reference distribution while QI requires to take the distribution of one of the two samples as 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 to detect a larger range of alternative hypothesis, such as the nesting situation mentioned above. We discuss the

critical values of $K(X,Y)$ in the following section.

2.4 Computing critical values

Deriving the asymptotic distribution of $K(X,Y)$ is nontrivial since $K(X,Y)$ explicitly depends on two non *iid* processes, $D(X_k,P_n)$ and $D(Y_k,Q_m)$. This renders standard results on the Kolmogorov-Smirnov test inapplicable. Nevertheless, we posit without formal proof that $K(X,Y)$ follows the same limiting distribution as the regular Kolmogorov-Smirnov two sample statistic, i.e.

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

where

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

Although we are unable to prove this general results, we consider two special cases below and both are shown to conform to the conjecture. Our extensive simulation studies also demonstrate this general convergence.

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 such case $\widehat{F}_n(X_k)$ in (3) becomes a random variable t $\sim$uniform$[0,1]$ and thus $\widehat{G}_m(X_k)$ in (4) becomes $\widehat{G}_m(t)$. Then $K_{P_n}(X,Y)$ in (5), which is the test statistics for this special case, reduces to

$$K_P(t) = \sup_{t\in[0,1]} |t - \widehat{G}_m(t)|.$$

Since $\widehat{G}_m(t)$ is an empirical distribution of the i.i.d random variables $\{D(Y_1,P),\ldots,D(Y_m,P)\}$, $K_P(t)$ is exactly the one sample Kolmogorov-Smirnov statistic for testing the uniformity of $\widehat{G}_m(t)$. Therefore,

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

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 $K_{Q_m}(X, T)$) converges to 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 special cases to more general setting is challenging. This issue was noted in **?** where the authors conjectured that their two sample QI asymptotically followed a normal distribution, as its one sample version. Their conjecture was only later proven in **?** after substantial theoretical development. The techniques that emerged from the proof in **?** relied heavily on QI being an expectation, making them largely inapplicable to our context that involves suprema. In lieu of an asymptotic distribution we may consider using permutation to find critical values for $K(X, Y)$ **[Bo: add a reference for purmutation]**. Permutation works well for small samples or sparsely observed functions, however 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 $K(X, Y)$, and size and power of the test. Each of these properties is evaluated using two dimensional functional data since our main application considers ensembles of spatial fields. All functional data in the simulation are generated from Gaussian random processes with an exponential covariance function $C(x, x') = \exp\{-\|x - x'\|/r\}$, where the range parameter r governs how quickly the corre-

lation decays between observations. A small (larger) r indicates a weak (strong) correlation and consequently a rough (smooth) functional data. We could instead use Matérn covariance function **[Bo: add reference]** to control the smoothness **[Bo: by looking at the definition of smoothness in http://anson.ucdavis.edu/ mueller/Review151106.pdf and the mean square differentiability in my spatial class notes, it looks Matern may be more appropriate?]** In each simulation we consider the sample X as the baseline and Y as the sample to be varied.

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 is approximated by the Kolmogorov distribution, even at moderate sample size. The functional data X and Y are generated with mean $\mu = 0$ and standard deviation $\sigma = 1$ on spatial domain $[0, 20] \times [0, 20]$. Since the integrated Tukey depth is invariant to location and scale, we only vary the range parameter r to be 5, 10, 15, and 20 as well as the number of replicates n to be 25, 50, 100 and 150 in studying the convergence of the asymptotic distribution of the test statistics. Unbalancing sample sizes enabled much faster convergence since $K(X, Y)$ would behave more like either of its one sample versions **[Bo: Is this shown in Figure 2 and 3?]**. The permutation distribution was constructed by recomputing $K(X, Y)$ on 2000 permutations of the generated X and Y samples. Then we calculate the L^2 distance between the permutation distribution and the Kolmogorov distribution. In addition, we also calculate the difference in critical values derived from either the Kolmogorov or the permutation distribution at three common significance levels: 0.01, 0.05 and 0.10. We run simulation 100 times for each combination of r and n to obtain the boxplots in Figures 2 and 3.

Figure 2 demonstrates convergence of the permutation distribution to Kolmogorov in L^2. For even small sample sizes such as $n = 25$, the distance between the two distributions is

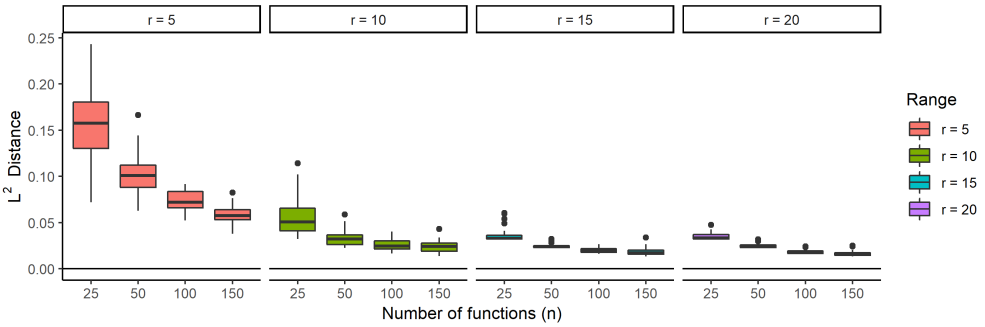

Figure 2: L^2 distance between the permutation distribution and the Kolmogorov distribution under 16 different range and sample size settings.

already vanishingly small for smooth data ($r = 10, 15, 20$). The $r \leq 5$ case is typically not an obstacle in practice because functional data would usually be preprocessed with a smoothing step. In all cases the L^2 distance decays rapidly with an increasing sample size so even noisy data can be compensated for if the sample size is large enough.

Figure 3 evaluates the convergence of two sets of critical values at three common significance levels: 0.01, 0.05 and 0.10. This figure aims to answer the question of how much bias if any in making decisions with the asymptotic Kolmogorov distribution and the permutation, even if the two distributions are not in exact agreement. Again, a sufficient amount of smoothness ($r > 5$) is required to have well behaved critical values. If the data is not sufficiently smooth the Kolmogorov distribution tends to have smaller critical values than its 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 $r \geq 10$, the critical values of the permutation distribution become highly agreeable to the Kolmogorov's. The observed differences are minuscule so any decision reached using the Kolmogorov distribution is likely to be same as if the permutation distribution were used.

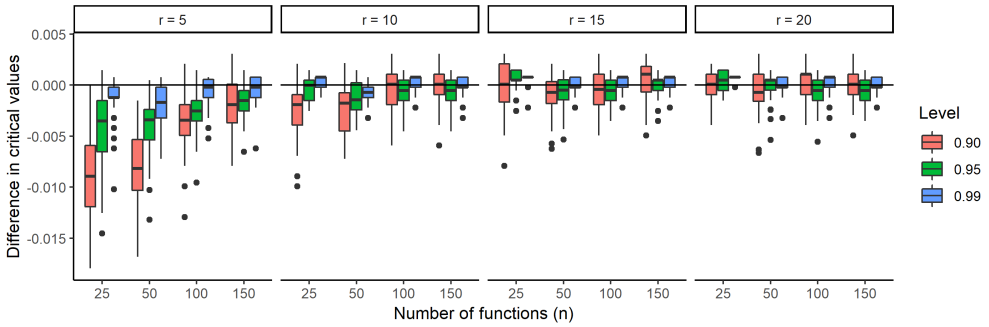

Figure 3: Kolmogorov critical values minus permutation critical values at three common test levels: 0.90, 0.95. 0.99 under 16 different range and sample size settings.

3.2 Size and power

Using the same data generating process as in 3.1, we evaluate the size of our test using critical values from the asymptotic Kolmogorov distribution and also compare our size to the QI test. Again only the range r and sample size n will be varied. The size under each combination of r and n was estimated 50 times **[Bo: How to get size 50 times for each simulation?]** using 1000 simulations apiece; the results of which are summarized in Figure 4. Our simulations show that for even a small sample such as $n = 50$, our test is able to control the size near the prescribed level if $r > 5$. As in the convergence simulation this smoothness condition is not all that impactful in practice since functions are typically smoothed before analysis. For functional data with $r \geq 10$, the size of our test is controlled very near the nominal level even for small sample sizes. Moreover, the range no longer seems to play a role beyond a threshold between 5 and 10 for the spatial domain $[0, 20] \times [0, 20]$. The QI test appears to inflate the size in all cases compared to our test.

Then we compare the power of our test and QI test in detecting changes in the three parameters μ, σ, and r which govern the underlying Gaussian process in our data generation. For power calculation, the sample size was fixed at $n = m = 400$ functions per sample so that K and QI would have comparable type I error rates**[Bo: why?]**. The functions in X are still generated from the Gaussian process with $\mu = 0$ and $\sigma = 1$ while r set to four different values: 5, 10, 15 and 20. This yields four baseline models for X at different smoothness

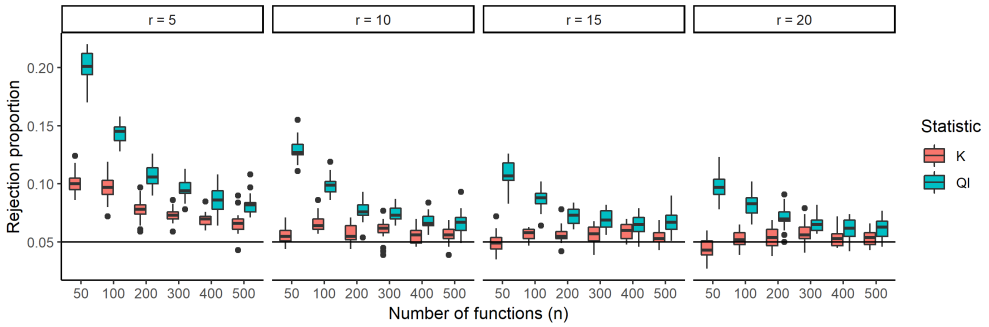

Figure 4: Size for K and QI under 16 different combinations of range and sample size ordered from noisiest (r = 5) to smoothest (r = 20). Black line at 0.05 designates the nominal level of the test.

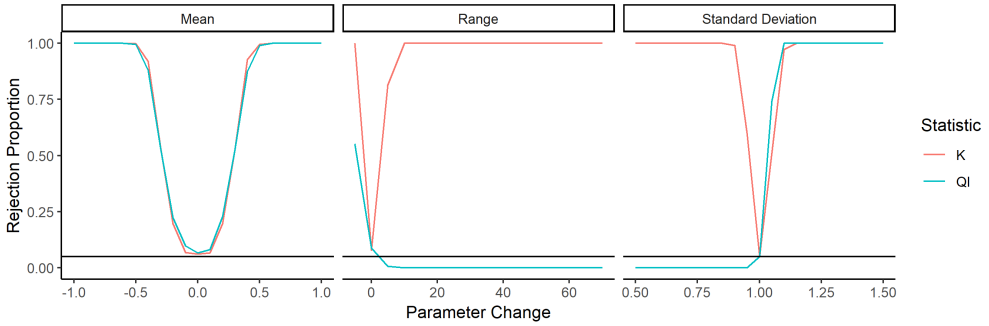

Figure 5: Power of K and QI in detecting changes in the three parameters in Gaussian process. Mean and Range are presented as shifts of parameters in Y from X. Standard Deviation is presented as a multiple of standard deviation in X.

levels. We then generate Y samples corresponding to each baseline model by setting the parameter of interest to different values. In order to examine the power curve we vary the mean of Y from -1 to 1, the standard deviation from 0.5 to 1.5 and the range from 5 to 80. We compute the power of our test to detect changes from each of these four baseline models in all three parameters, and then compare to the QI test. Each power is calculated from 500 simulation runs **[Bo: why the number of simulation runs keeps changing?]**

Figure 5 shows power functions for each parameter when X has $r = 10$. Full results can be found in the Appendix. Essentially, increasing the smoothness only flattens the power function slightly. Both $K(X, Y)$ and QI are almost equally powerful in detecting changes in mean, decreases in range, and increases in standard deviation. However, $K(X, Y)$ shows

improvement over QI in detecting increases in range or decreases in standard deviation. It should be noted that QI was explicitly designed to ignore decreases in standard deviation, since in their application a drop in standard deviation was desirable.

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 is kept constant throughout at $r = 20$. 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 were generated as follows:

$$\mu(s) = t \sin\left(2\pi \frac{s_1}{20}\right) \otimes t \sin\left(2\pi \frac{s_2}{20}\right)$$
$$\sigma(s) = t \sin\left(2\pi \frac{s_1}{20}\right) \otimes t \sin\left(2\pi \frac{s_2}{20}\right) + 1,$$

[Bo: why there are four equations?]

$$\mu(s) = \left(0.8t \cos\left(\pi \frac{s_1}{20}\right) + 1\right) \otimes \left(0.8t \cos\left(\pi \frac{s_1}{20}\right) + 1\right) - 1$$
$$\sigma(s) = \left(0.8t \cos\left(\pi \frac{s_1}{20}\right) + 1\right) \otimes \left(0.8t \cos\left(\pi \frac{s_1}{20}\right) + 1\right),$$

where $s = (s_1, s_2)$ and t was varied from 0.05 to 1 in increments of 0.05. We fixed the sample size to $n = 500$ and used 1000 simulations per t value to estimate the power at t.

Figure 6 shows the power functions of K and QI under heterogeneous mean and standard deviation changes. For detecting mean changes, both K and QI maintain comparable powers although our test indeed carries more power than the QI test at certain range of mean change.

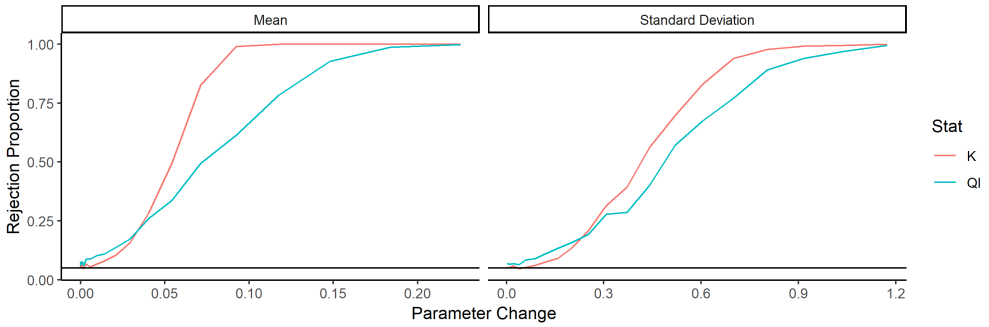

Figure 6: Power functions for K and QI under heterogeneous differences in the mean and standard deviation between X and Y. Powers were plotted against the L^2 distance in the mean and standard deviation between X and Y.

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. However, huge difference between K and QI is observed when the standard deviation change is heterogeneous. Basically, K still completely maintains its power while QI fails.

4 Application to Data Assimilation

We used the proposed test K to detect differences in the background and analysis states of the 1000 year two meter surface temperature reconstruction. Differences at the global level are investigated in section 4.1 using the full spatial extent of the background and analysis states. In section 4.2 the background and analysis are partitioned into 12 regions corresponding to the 5 oceans and 7 continents. We investigate how differences at the global level distribute down to these regions and how correlation between regions may impact K.

4.1 Global Reconstructions

Figure 7 shows the magnitude of K over time along with the associated p-values. The p values were adjusted by the Benjamini-Hochberg procedure to have a false discovery rate of 0.05.

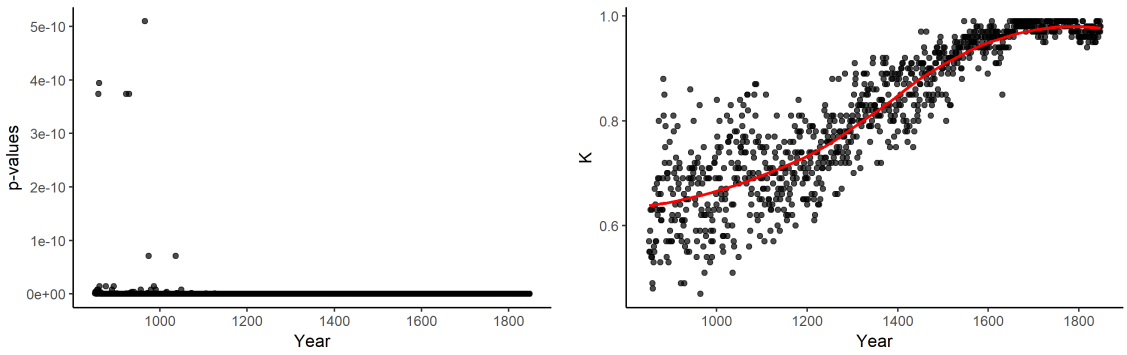

Figure 7: Value of K (left) and associated p-values (right) over the reconstruction years 850 CE to 1850 CE. Larger values of K indicate large differences in distribution.

We interpret their near uniformity about zero as a strong indication that the background and analysis are different in distribution each year. Consequently we believe that the proxies are indeed changing the distribution of the background and hence having a material influence over the assimilation reconstructions.

The magnitudes of K indicate an initial moderate separation between the background and analysis that steadily increases over time until the end of the reconstruction period. The apparent rise in separation is explained by the fact that proxy information is also steadily introduced into the model over time. As the reconstruction nears present day more proxies become available for assimilation and consequently the data assimilation fields should diverge further from the background.

4.2 Regional Variations

Analysis of the global fields is important for establishing the strength of proxy influence in the model as a whole and for confirming the existence of its upward trend. A natural next step then is to consider how these effects distribute down to a regional level; namely how proxies impact climatological estimation at the continental and oceanic level. Proxies are not collected uniformly across all regions and there is a clear over representation in North America and Europe and under representation in Africa, South America and large swathes

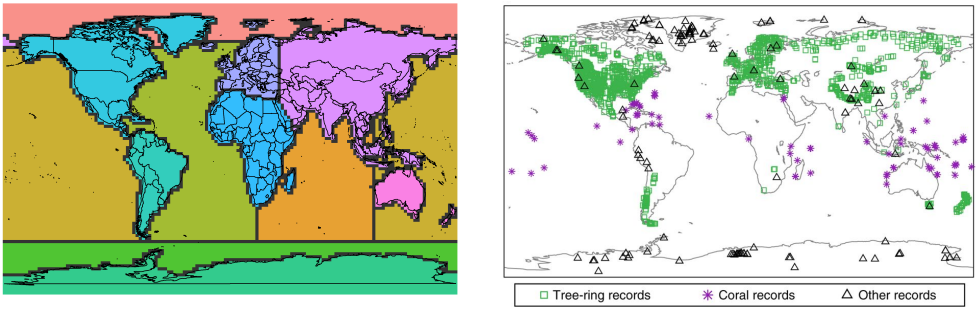

Figure 8: Regions versus proxy locations. The vast majority of proxies are collected in convenient locations such as North America and Europe. Not all of the displayed proxies are available every year in the reconstruction, as reconstruction gets closer to the present day more become available.

of the oceans, Fig. 8. We would thus not expect proxies to have as strong of an effect over these latter regions as we would the former. Still, owing to a long range correlation structure evidenced between surface temperature and the proxies (Fig. 10) we believe that these regional deficits may be somewhat mitigated. Our results provide a measure of support for this hypothesis since we see strong divergence between the background and analysis in Africa and the Pacific.

We split the global background and analysis ensembles into 12 regions corresponding to the five oceans and seven continents, Fig. 8. Within each of the twelve regions we used our K statistic to test for differences in the background and analysis states over the full reconstruction period. Our findings are summarized in Figure 9 and supported by Figures 10 and 11. Figure 9 shows K's progression over time for each region, analogous to the global study in Fig. 4.1. The increasing effect size over time in each region is determined most strongly by the increasing proxy information availability in time. As proxy information is gradually introduced over time the analysis states becomes more and more distinct from the background. This effect holds true even for some regions where proxy information is relatively scarce. However, because the strength of the climatological connections between various regions are different from one another, the proxy information will not be uniformly dispersed and so not all regions will benefit equally. Therefore it can be helpful to consider

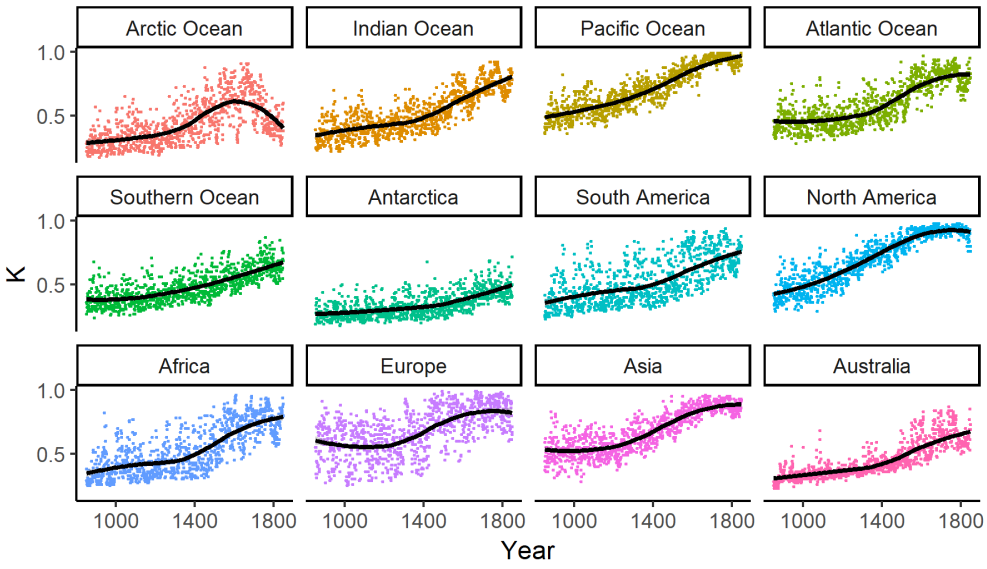

Figure 9: K over time by region. Regional K values were computed by only measuring differences in the region of interest within the Global reconstructions. They generally follow the pattern of the Global K values with the exception of the Arctic Ocean.

the spatial extent over which proxies in different locations can have an influence on the analysis state. This can be estimated by looking at proxy-point correlation maps at different points in time, Fig. 10. A thorough explanation of how these maps are created can be found in the appendix.

The maximum r^2 values decrease further back in time as fewer proxies are available. The spatial extent of the correlations thus provide a guide to interpreting the general decrease in effect size as well as the differences in the effect size between regions. Regions with

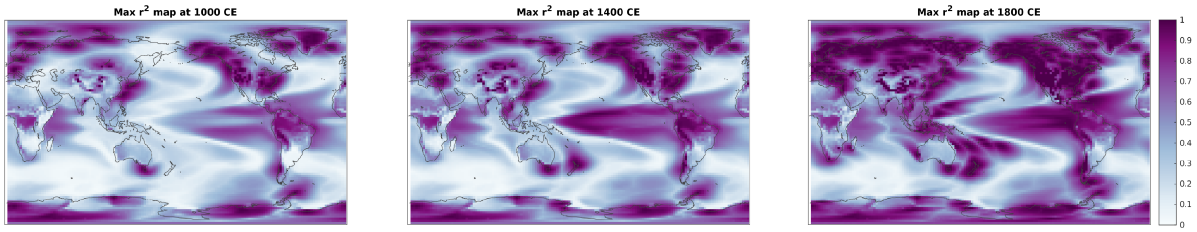

Figure 10: Proxy-point correlation maps for representative years 1000, 1400, and 1850 CE. There is an overall increasing proxy point correlation (purple) in time. This is reflected in the increasing effect sizes seen in regions with little proxy representation.

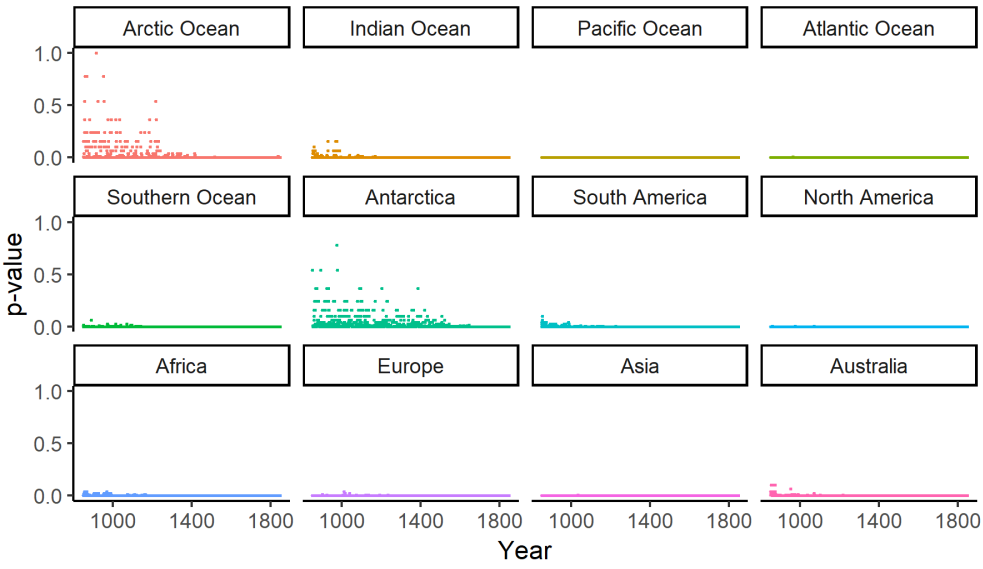

Figure 11: P-values of K over time by region

higher max r^2 values generally have higher effect sizes (e.g., comparing North America with Australia). Note that some regions (particularly the Pacific ocean) can have few proxies but are highly correlated with other regions of the globe that have many proxies (seen in the larger max r^2 values in Fig. 10), therefore their effect sizes can remain high despite the lack of local proxies.

The Arctic and Southern oceans represent two anomalies with regards to their supposed proxy information. Most noticeably around 1600 the Arctic ocean experiences a strong trend reversal in K, just when other regions are experiencing trend increases. This runs counter to the fact that number proxies and the proxy-point correlation are both increasing in the Arctic over this time period. Conversely the Southern ocean has relatively large values of K when proxy information would lead us to believe they should be much smaller. The Southern ocean has no collected proxies and it has some of the weakest overall proxy-point correlation strength, yet it experiences a strong and significant divergence between its background and analysis states. It is currently unclear why either the Arctic or Southern ocean are behaving differently than their levels of proxy information would indicate.

5 Discussion

By estimating differences in the background and analysis's distributions we were able to investigate the influence proxies hold over the data assimilation process. Our investigation centered on studying the existence and degree of influence, how influence develops over time, and how influence is distributed spatially. Existing two sample functional data tests were found to be insufficient for answering these questions due to both testing and data limitations. Most tests only consider specific quantities such as mean differences which do not represent the full role that proxies can play. Of those that do consider distributional differences they either rely on bootstrapping (hall) and are thus too slow for data of this size or they utilize basis representations (strecu) which struggle to capture the subtle variations proxies induce. To overcome these limitations we developed a new non-parametric two sample test for functional distributions. This test was seen to control the size well, even with small sample sizes, and to be powerful against changes in location, scale, and correlation structure. Its foundation in a computationally efficient data depth function means that it is also fast enough to evaluate on large climate data sets, such as considered here.

Our results on the global reconstructions provide strong evidence of a clear separation between the background and analysis, but that the degree of separation depends heavily on reconstruction location and period. There was seen to be overall upward trend in proxy influence which is generally maintained even when subdividing the data 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 is the first rigorous confirmation of the long standing, educated, belief that increasing proxy information should correlate with a commensurate increase in separation.

It was also seen that, despite the stark imbalance in proxy density across the various regions, most regions still exhibited an increasing separation. This mitigating effect is mostly attributable to the long range correlation structure proxies and temperatures often display.

Some regions such as Pacific and South America have very few local proxies but due to their significant correlation with other regions still benefit from proxies collected remotely.

Looking forward, our results indicate that as more proxy data is collected and assimilated climate models will increasingly reflect the climatic states of past Earth. This has far reaching consequences for those who use long run paleoclimate reconstructions to inform predictions about future climate. Furthermore the two sample test developed here is much more broadly applicable than for studying proxy influence. Our generic formulation allows it to be applied seamlessly on any functional data that the depth function can handle, including curves in $\mathbb{R}$ and higher dimensional functions in $\mathbb{R}^n$. We hope that future work can both establish our tests asymptotic distribution and study its efficacy in higher dimensions.

Also, because our test is based on integrated depth and not distances or principal components the assumptions we need to make about the data are very light. We do not need to assume that the curves are square integrable, second order stationary, densely sampled, or even strictly continuous. Computing the integrated Tukey depth merely requires that the functions are almost everywhere continuous and observed at the same locations. The first requirement allows us to consider discrete or continuous functions without any modification to the procedure. The second is typically not an issue for densely sampled data since it can be interpolated onto a shared grid with little loss in accuracy. The assumptions that P and Q are absolutely continuous distributions and that each X_i and Y_j is univariate could also be relaxed. Letting P and Q be discrete distributions simply changes the measure used for integration to the counting measure. Each X_i and Y_j could also be multivariate valued so long as they both map to the same subspace of $\mathbb{R}^p$. This only changes our integration to be over a multivariate depth instead of a univariate depth. We did not explore or test these generalized settings in this paper but we make note of them to highlight the flexibility allowed for by depth based testing.

Acknowledgments

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. Working datasets were provided by Jason Smerdon and Nathan Steiger of the Lamont-Doherty Earth Observatory (LDEO) at Columbia University.

A Appendix

A.1 Proofs

Proposition 2.1

Proof. let P be a distribution on $C[0,1]^p$ and suppose $X = \{X_1, ..., X_n\}$ and $Y = \{Y_1, ..., Y_n\}$ are two i.i.d 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

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

Since $n \gg m$. 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. □

A.2 Correlation Maps

The correlation maps in figure 10 are constructed by first computing point correlation maps for the background 2 m temperature time series for each proxy location: the background is a continuous climate model simulation and the correlation is computed between the 2 m temperature time series at a given proxy location and all global grid point 2 m temperature time series in the background. This generates nearly 3,000 correlation maps, one for each proxy location. Then for the representative years of 1000, 1400, and 1800 CE, the maximum r^2 value is found for each grid point among all the correlation maps that correspond to the proxies that are available during those specific years.

A.3 Power Plots