Research Article Environmetrics

Received: 19 November 2015, Revised: 19 April 2016, Accepted: 22 April 2016, Published online in Wiley Online Library: 23 May 2016

(wileyonlinelibrary.com) DOI: 10.1002/env.2395

Comparison between spatio-temporal random processes and application to climate model data

Bo Li[a]*, Xianyang Zhang[b] and Jason E. Smerdon[c]

Comparing two spatio-temporal processes are often a desirable exercise. For example, assessments of the difference between various climate models may involve the comparisons of the synthetic climate random fields generated as simulations from each model. We develop rigorous methods to compare two spatio-temporal random processes both in terms of moments and in terms of temporal trend, using the functional data analysis approach. A highlight of our method is that we can compare the trend surfaces between two random processes, which are motivated by evaluating the skill of synthetic climate from climate models in terms of capturing the pronounced upward trend of real-observational data. We perform simulations to evaluate our methods and then apply the methods to compare different climate models as well as to evaluate the synthetic temperature fields from model simulations, with respect to observed temperature fields. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: comparing spatiotemporal processes; functional data analysis; covariance operator; mean surface; temporal trend

1. INTRODUCTION

The comparison between spatio-temporal random processes is often desired for addressing problems in atmospheric science, climatology, and the environmental sciences more generally. For example, general circulation models (GCMs) have been widely used as global climate models to simulate synthetic climates that are interpreted as reasonable emulations of the real-climate system. An important element of model assessments is therefore to compare different GCMs as well as to evaluate their performance in comparison with the real climate. The assessment of the difference between GCMs requires us to compare the synthetic climate generated from different climate models, while the evaluation of their performance requires us to compare model-simulated climatic features to the real climate. Another example is that to quantify the difference between the methods used for reconstructing spatio-temporal fields of past climates from climate proxies such as tree-rings, ice cores, corals, and so on, a natural exercise is to compare the reconstructed climate fields that are derived from each method. All these examples call for comprehensive means of comparing spatio-temporal random processes, which nevertheless, presents challenges because of the high-dimensional characteristic and the typically complex dependency structure in spatio-temporal random fields.

To date, the literature on comparing two random fields primarily focuses on comparing either two spatial processes or two time series. Briggs and Levine (1997), Shen *et al.* (2002), and Pavlicová *et al.* (2008) compared two spatial random fields over grids based on the wavelet transform. Diebold and Mariano (1995) evaluated whether two competing forecasts in time series are equally accurate by examining whether the loss function defined based on prediction errors is significantly different from zero. Later, Snell *et al.* (2000) and Wang *et al.* (2007) extended Diebold and Mariano (1995) to the context of spatial data. Hering and Genton (2011) developed a hypothesis testing to evaluate the difference between two spatial random fields in terms of user-chosen loss functions, and the innovation of their method lies in considering the spatial correlation of the loss differential in estimating its uncertainty. Lund and Li (2009) for the first time evaluated the difference between two time series by jointly examining the discrepancies in their mean and covariance structure.

The comparison between two spatio-temporal random fields is mainly conducted from either geostatistics or the functional data analysis point of view. Enlightened by Lund and Li (2009), Li and Smerdon (2012) attempted to compare climate field reconstructions by integrating the difference in both the mean and covariance structure based on the two-sample Kolmogorov–Smirnov test of whitened random fields. Their method nevertheless is unsatisfactory in several ways. Firstly, their approach treated each time-specific spatial field separately, although it would be more advantageous to take their temporal dependence into account. Secondly, their testing method is sensitive to the misspecification of the parametric covariance model chosen for the random field. And lastly, their approach assumed that the two spatio-temporal random fields are independent from each other, yet this is not always the case in some testing applications that share underlying data. The functional data analysis approaches include the tests for the equality of mean functions (Fan and Lin, 1998; Cuevas *et al.*, 2004; Horváth *et*

* *Correspondence to: Bo Li, University of Illinois at Urbana-Champaign, Champaign, IL 61820, U.S.A. E-mail: libo@illinois.edu*

a *Department of Statistics, University of Illinois at Urbana-Champaign, 725 S Wright St., Champaign, IL 61820, U.S.A.*

b *Department of Statistics, Texas A&M University, College Station, College Station, TX 77840, U.S.A.*

c *Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY 10964, U.S.A.*

al., 2013), and the test for the equality of the second-order structure (Benko *et al.*, 2009; Panaretos *et al.*, 2010; Fremdt *et al.*, 2013; Kraus and Panaretos, 2012; Zhang and Shao, 2015).

All the aforementioned functional methods only test either the mean or covariance but not the two jointly. The methods also all assume independence between the compared random fields except for Zhang and Shao (2015), which has several nice properties including the consideration of temporal dependence and dependence between two functional datasets and the additional advantage of being non-parametric compared with Li and Smerdon (2012). Motivated by the scientific interests of evaluating the comprehensive difference between global climate models and the skill of each climate model in mimicking the real climate, in particular, their striking upward trend, we therefore propose testing methods in the framework of Zhang and Shao (2015) to compare various characteristics between two spatio-temporal random fields. Our proposed method inherits all the merits in Zhang and Shao (2015) and furthermore features the ability to jointly evaluate the first and second moments of random fields and to evaluate the trend surfaces of the random fields. We will apply our method to synthetic climate data generated from climate models and observed climate data.

The paper is organized as follows. Section 2 first reviews the method in Zhang and Shao (2015) and then proposes new statistical methods for assessing the difference between two random fields in terms of mean surface and covariance structure, and the linear trend surface. Section 3 reports simulation results to evaluate the sizes and powers of the proposed methods. Section 4 applies the proposed methods to assess global climate models. Section 5 summarizes our methods and discusses the directions for further investigation, and Appendix section demonstrates the proof for the theorem developed in this paper.

2. COMPARISON METHODS

Two hypothesis testings for comparing spatio-temporal random fields are developed. One extends the methods in Zhang and Shao (2015) to jointly evaluate the mean and covariance structure, while the other evaluates the difference of temporal trends between two random fields. The first method is motivated by offering a single measure of the difference between two climate models, and the second is motivated by assessing their skill in capturing the pronounced upward trend of the recent climate change.

Let $X(\mathbf{s},t)$ and $Y(\mathbf{s},t)$ be two spatio-temporal random fields observed over spatial locations, $\mathbf{s}\in D$, and time points, $t\in\mathbb{Z}$. Because we treat them here as two temporally dependent functional processes, we change their notation to $\{X_t(\mathbf{s})\}_{t=1}^{N_1}$ and $\{Y_t(\mathbf{s})\}_{t=1}^{N_2}$ to better reflect that they are considered as functional time series. Let $\mathbb{H}$ be the Hilbert space of square integrable functions over $D\subseteq\mathbb{R}^2$. For any functions $f,g\in\mathbb{H}$, the inner product between f and g is defined as $<f,g>=\int_D f(\mathbf{s})g(\mathbf{s})d\mathbf{s}$, and $||f||=<f,f>^{1/2}$ denotes the inner product induced norm. Define the operator $f\otimes g(\cdot)=<f,\cdot>g$ for $f,g\in\mathbb{H}$ such that for a function h, the operator $f\otimes g(h)=<f,h>g$ maps h to $<f,h>g$. Let $L_{\mathbb{H}}^p$ be the space of $\mathbb{H}$-valued random variables X such that $E||X||^p<\infty$ for some $p>0$.

2.1. Testing mean and covariance function

Assuming that the functional time series are time invariant, that is, the spatio-temporal random fields are second-order stationary in time, we define $\mu_X(\mathbf{s})=E\{X_t(\mathbf{s})\}$ and $\mu_Y(\mathbf{s})=E\{Y_t(\mathbf{s})\}$ as their mean functions over $\mathbf{s}\in D$. Furthermore, we define $C_X(\mathbf{s},\mathbf{s}')=E[\{X_t(\mathbf{s})-\mu_X(\mathbf{s})\}\{X_t(\mathbf{s}')-\mu_X(\mathbf{s}')\}]$ and $C_Y(\mathbf{s},\mathbf{s}')=E[\{Y_t(\mathbf{s})-\mu_Y(\mathbf{s})\}\{Y_t(\mathbf{s}')-\mu_Y(\mathbf{s}')\}]$ as the covariance functions of $X_t(\mathbf{s})$ and $Y_t(\mathbf{s}')$ over $\mathbf{s},\mathbf{s}'\in D$, respectively. Depending on the problem of interest, three types of hypotheses on mean and covariance functions can be formulated as follows:

I	$H_0:\mu_X=\mu_Y$;	$H_a:\mu_X\neq\mu_Y$;
II	$H_0:C_X=C_Y$;	$H_a:C_X\neq C_Y$;
III	$H_0:\mu_X=\mu_Y$ & $C_X=C_Y$;	$H_a:\mu_X\neq\mu_Y$ or $C_X\neq C_Y$.

Here, "$\mu_X\neq\mu_Y$" and "$C_X\neq C_Y$" mean that $||\mu_X-\mu_Y||>0$ and $\int_{D\times D}|C_X(\mathbf{s},\mathbf{s}')-C_Y(\mathbf{s},\mathbf{s}')|^2d\mathbf{s}d\mathbf{s}'>0$, respectively. In particular, Hypothesis III arises out of the interest in providing a comprehensive assessment between two climate fields in terms of the first and second moment structures. Hypotheses I and II have been addressed by Zhang and Shao (2015), so here, we focus on Hypothesis III. Because our method for Hypothesis III is an extension of the methods for Hypotheses I and II, we will first give a brief review of the techniques for performing the first two hypotheses and then present our test statistic.

2.1.1. *Review of methods in Zhang and Shao (2015)*

Zhang and Shao (2015) developed their testing methods based on functional principal component analysis and the recently developed self-normalization technique. Let $N=N_1+N_2$ be the total time length for two random fields. Define the recursive sample mean functions $\widehat{\boldsymbol{\mu}}_{X,m}=\frac{1}{m}\sum_{t=1}^m X_t$, and $\widehat{\boldsymbol{\mu}}_{Y,n}=\frac{1}{n}\sum_{t=1}^n Y_t$ with $1\leqslant m\leqslant N_1$ and $1\leqslant n\leqslant N_2$, and the pooled sample covariance operator

$$\widehat{C}_{XY}=\frac{1}{N}\left[\sum_{t=1}^{N_1}\{X_t-\widehat{\boldsymbol{\mu}}_{X,N_1}\}\otimes\{X_t-\widehat{\boldsymbol{\mu}}_{X,N_1}\}+\sum_{t=1}^{N_2}\{Y_t-\widehat{\boldsymbol{\mu}}_{Y,N_2}\}\otimes\{Y_t-\widehat{\boldsymbol{\mu}}_{Y,N_2}\}\right]$$

The eigenvalues and eigenfunctions corresponding to $\widehat{C}_{XY}$ are denoted by $\{\widehat{\lambda}_{XY}^j\}$ and $\{\widehat{\phi}_{XY}^j\}$. Then, define a sequence of vectors consisting of the projected (recursive) mean differences on the first K eigenfunctions:

$$\widehat{\boldsymbol{\psi}}_k=\left(<\widehat{\boldsymbol{\mu}}_{X,\lfloor kN_1/N\rfloor}-\widehat{\boldsymbol{\mu}}_{Y,\lfloor kN_2/N\rfloor},\widehat{\phi}_{XY}^1>,\ldots,<\widehat{\boldsymbol{\mu}}_{X,\lfloor kN_1/N\rfloor}-\widehat{\boldsymbol{\mu}}_{Y,\lfloor kN_2/N\rfloor},\widehat{\phi}_{XY}^K>\right)^{\mathrm{T}}$$

for $2 \leqslant k \leqslant N$, where $\lfloor W \rfloor$ is the largest integer not greater than $W \in \mathbb{R}$. The test statistic for Hypothesis I is

$$TS1(K) = N\widehat{\boldsymbol{\psi}}_N^{\mathrm{T}}\mathbf{V}_{\psi}^{-1}(K)\widehat{\boldsymbol{\psi}}_N$$

where $\mathbf{V}_{\psi}(K) = \frac{1}{N^2}\sum_{k=1}^{N} k^2(\widehat{\boldsymbol{\psi}}_k - \widehat{\boldsymbol{\psi}}_N)(\widehat{\boldsymbol{\psi}}_k - \widehat{\boldsymbol{\psi}}_N)^{\mathrm{T}}$. The parameter K is a user-chosen number that determines the number of eigenfunctions to be used in the test. It is also associated with the cumulative percentage of total variation with respect to the pooled sample covariance.

The K-length vector $\widehat{\boldsymbol{\psi}}_k$ consists of projected differences, with the jth element being projected onto the jth eigenfunction $\widehat{\phi}_{XY}^{j}$. The index k for $\widehat{\boldsymbol{\psi}}_k$ indicates the kth paired difference between the recursive estimates of mean functions. Because of these recursive estimates that are the kernel of the self-normalization technique, we allow each individual data to be temporally correlated and moreover the two data sets to be correlated if additionally assuming $N_1/N_2 \to 1$.

Similarly, to test the covariance function, we define the recursive covariance estimators $\widehat{C}_{X,m} = \frac{1}{m}\sum_{t=1}^{m}\{X_t - \widehat{\boldsymbol{\mu}}_{X,N_1}\}\otimes\{X_t - \widehat{\boldsymbol{\mu}}_{X,N_1}\}$ and $\widehat{C}_{Y,n} = \frac{1}{n}\sum_{t=1}^{n}\{Y_t - \widehat{\boldsymbol{\mu}}_{Y,N_2}\}\otimes\{Y_t - \widehat{\boldsymbol{\mu}}_{Y,N_2}\}$ with $1 \leqslant m \leqslant N_1$ and $1 \leqslant n \leqslant N_2$. Then, we define a sequence of matrices formed by the projected covariance differences, $\mathbf{C}_k = [c_k^{i,j}]$, where

$$c_k^{i,j} = < \left(\widehat{C}_{X,\lfloor kN_1/N\rfloor} - \widehat{C}_{Y,\lfloor kN_2/N\rfloor}\right)\widehat{\phi}_{XY}^{i}, \widehat{\phi}_{XY}^{j} >,\ 2 \leqslant k \leqslant N,\ 1 \leqslant i, j \leqslant K$$

Let $\widehat{\boldsymbol{\alpha}}_k$ be the vectorized $\mathbf{C}_k$, which contains the elements on and below the main diagonal of $\mathbf{C}_k$ (i.e., vectorizing only the lower triangular part). The test statistic for Hypothesis II is

$$TS2(d) = N\widehat{\boldsymbol{\alpha}}_N^{\mathrm{T}}\mathbf{V}_{\alpha}^{-1}(d)\widehat{\boldsymbol{\alpha}}_N$$

where $d = K(K+1)/2$ and $\mathbf{V}_{\alpha}(d) = \frac{1}{N^2}\sum_{k=1}^{N} k^2(\widehat{\boldsymbol{\alpha}}_k - \widehat{\boldsymbol{\alpha}}_N)(\widehat{\boldsymbol{\alpha}}_k - \widehat{\boldsymbol{\alpha}}_N)^{\mathrm{T}}$. Again, K is a user-chosen number and represents the cumulative percentage of total variation.

The pivotal limiting distributions of $TS1(K)$ and $TS2(d)$ are derived in Zhang and Shao (2015). Define $\mathbf{B}_q(r)$ as a $q-$dimensional vector of independent standard Brownian motions. Let $W_q = \mathbf{B}_q(1)^{\mathrm{T}}\mathbf{J}_q^{-1}\mathbf{B}_q(1)$, where $\mathbf{J}_q = \int_0^1\{\mathbf{B}_q(r) - r\mathbf{B}_q(1)\}\{\mathbf{B}_q(r) - r\mathbf{B}_q(1)\}^{\mathrm{T}}dr$, then $TS1(K)$ and $TS2(d)$ converge to W_K and W_d, respectively. The empirical distributions of W_q for any q can be obtained numerically by approximating the standard Brownian motion with the standardized partial sum of i.i.d standard normal random variables.

2.1.2. Test for Hypothesis III

In addition to examining the mean and covariance function separately, it is also of great interest to assess the climate fields by jointly evaluating their first moment and second moment structures. Therefore, we propose a test statistic for Hypothesis III by integrating the test statistics for Hypotheses I and II.

Proposition 2.1 *Denote by* $\widehat{\boldsymbol{\beta}}_k = (\widehat{\boldsymbol{\psi}}_k^{\mathrm{T}}, \widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}})^{\mathrm{T}}$ *with* $2 \leqslant k \leqslant N$. *Let* $s = K(K+3)/2$ *and* $\boldsymbol{V}_{\beta}(s) = \frac{1}{N^2}\sum_{k=1}^{N} k^2(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_N)(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_N)^{\mathrm{T}}$. *Under Assumptions 6.1, 6.2, 6.4, and 6.5 in the Appendix, we have*

$$TS3(s) = N\widehat{\boldsymbol{\beta}}_N^{\mathrm{T}}\boldsymbol{V}_{\beta}^{-1}(s)\widehat{\boldsymbol{\beta}}_N$$

converges to W_s. *Furthermore, assume* $N_1/N_2 \to 1$, *then the conclusion also holds with Assumption 6.2 replaced by Assumption 6.3.*

Notice that by replacing Assumption 6.2 with 6.3, we allow the two time series to be dependent. The proof of Proposition 2.1 follows the arguments in demonstrating the limiting distributions of $TS1$ and $TS2$ in Zhang and Shao (2015). The details are omitted. The test statistics essentially examined the aggregated differences of the mean surface and covariance structure projected onto the directions of major variation.

2.2. Assessment of trends between two random fields

When the trend appears a major feature of the spatio-temporal random fields, it is often of interest to compare the trends between two random fields. In order to formulate this problem, we consider the following model:

$$\begin{aligned} X_t(\mathbf{s}) &= \mu_X(\mathbf{s}) + t\beta_X(\mathbf{s}) + \epsilon_{X,t}(\mathbf{s}), \quad t = 1, 2, \ldots, N_1, \\ Y_t(\mathbf{s}) &= \mu_Y(\mathbf{s}) + t\beta_Y(\mathbf{s}) + \epsilon_{Y,t}(\mathbf{s}), \quad t = 1, 2, \ldots, N_2 \end{aligned} \tag{2.1}$$

where $\epsilon_{X,t}$ and $\epsilon_{Y,t}$ are both mean zero and stationary in time. We use $\beta_X(\mathbf{s})$ and $\beta_Y(\mathbf{s})$ to represent the spatially varying (linear) trends in $X_t(\mathbf{s})$ and $Y_t(\mathbf{s})$, respectively. Hence, our hypothesis for comparing the trends between two spatio-temporal random fields, $X_t(\mathbf{s})$ and $Y_t(\mathbf{s})$, can be written into

IV $\quad H_0 : \beta_X = \beta_Y;\ H_a : \beta_X \neq \beta_Y$

where "$\beta_X \neq \beta_Y''$ means that $||\beta_X - \beta_Y|| > 0$. Because β_X and β_Y are defined over an infinite dimensional functional space, it is necessary to reduce the dimensionality in developing the test statistic for Hypothesis IV. With some abuse of notation, let $C_{XY} = \gamma_1 E\epsilon_{X,1}\otimes\epsilon_{X,1} +$

$\gamma_2 E\epsilon_{Y,1} \otimes \epsilon_{Y,1}$, where $N_1/N \to \gamma_1$ and $N_2/N \to \gamma_2$. Denote by $\{\phi_{XY}^j\}$ the eigenfunctions of C_{XY}. To reduce the dimensionality, we project (2.1) onto the space spanned by $\{\phi_{XY}^j\}_{j=1}^K$, that is,

$$< X_t, \phi_{XY}^j > =< \mu_X, \phi_{XY}^j > + t < \beta_X, \phi_{XY}^j > + < \epsilon_{X,t}, \phi_{XY}^j >,$$
$$< Y_t, \phi_{XY}^j > =< \mu_Y, \phi_{XY}^j > + t < \beta_Y, \phi_{XY}^j > + < \epsilon_{Y,t}, \phi_{XY}^j > .$$

Let $\breve{X}_t = (< X_t, \phi_{XY}^1 >, \ldots, < X_t, \phi_{XY}^K >)^{\mathrm{T}}$, $\breve{\epsilon}_{X,t} = (< \epsilon_{X,t}, \phi_{XY}^1 >, \ldots, < \epsilon_{X,t}, \phi_{XY}^K >)^{\mathrm{T}}$, $\breve{\mu}_X = (< \mu_X, \phi_{XY}^1 >, \ldots, < \mu_X, \phi_{XY}^K >)^{\mathrm{T}}$, and $\breve{\beta}_X = (< \beta_X, \phi_{XY}^1 >, \ldots, < \beta_X, \phi_{XY}^K >)^{\mathrm{T}}$. These are the projections of each term in (2.1) on the first K joint eigenfunctions. These projections reduce the dimension of the data to K. Similarly, we can define $\breve{Y}_t, \breve{\epsilon}_{Y,t}, \breve{\boldsymbol{\mu}}_Y$, and $\breve{\beta}_Y$. Thus, we have

$$\breve{X}_t = \breve{\mu}_X + t\breve{\beta}_X + \breve{\epsilon}_{X,t},$$
$$\breve{Y}_t 7 = \breve{\mu}_Y + t\breve{\beta}_Y + \breve{\epsilon}_{Y,t}$$

In practice, C_{XY} is unknown and needs to be estimated from the data. Given preliminary estimators $\tilde{\beta}_X, \tilde{\beta}_Y, \tilde{\boldsymbol{\mu}}_X$, and $\tilde{\boldsymbol{\mu}}_Y$ and by letting $\tilde{\mathbf{m}}_X = \tilde{\boldsymbol{\mu}}_X + t\tilde{\beta}_X$ and $\tilde{\mathbf{m}}_Y = \tilde{\boldsymbol{\mu}}_Y + t\tilde{\beta}_Y$, we define the estimator for C_{XY} as

$$\widehat{C}_{XY} = \frac{1}{N}\left[\sum_{t=1}^{N_1}\{X_t - \tilde{\mathbf{m}}_X\} \otimes \{X_t - \tilde{\mathbf{m}}_X\} + \sum_{t=1}^{N_2}\{Y_t - \tilde{\mathbf{m}}_Y\} \otimes \{Y_t - \tilde{\mathbf{m}}_Y\}\right]$$

Let $\{\widehat{\phi}_{XY}^j\}$ be the eigenfunctions of $\widehat{C}_{XY}$. By projecting (2.1) onto the space spanned by $\{\widehat{\phi}_{XY}^j\}_{j=1}^K$, we have

$$< X_t, \widehat{\phi}_{XY}^j > =< \mu_X, \widehat{\phi}_{XY}^j > + t < \beta_X, \widehat{\phi}_{XY}^j > + < \epsilon_{X,t}, \widehat{\phi}_{XY}^j >,$$
$$< Y_t, \widehat{\phi}_{XY}^j > =< \mu_Y, \widehat{\phi}_{XY}^j > + t < \beta_Y, \widehat{\phi}_{XY}^j > + < \epsilon_{Y,t}, \widehat{\phi}_{XY}^j >$$

where $j = 1, 2, \ldots, K$. Define $\theta_X = (\breve{\boldsymbol{\mu}}_X^{\mathrm{T}}, \breve{\beta}_X^{\mathrm{T}})^{\mathrm{T}}$ and $\theta_Y = (\breve{\boldsymbol{\mu}}_Y^{\mathrm{T}}, \breve{\beta}_Y^{\mathrm{T}})^{\mathrm{T}}$. Let $Z_{t,K} = (\mathbf{I}_K, t\mathbf{I}_K)^{\mathrm{T}} \in \mathbb{R}^{2K \times K}$, where $\mathbf{I}_K$ denotes the $K \times K$ identity matrix. The recursive (least squares) estimator for θ_X is given by

$$\widehat{\theta}_{X,k} = \left(\widehat{\mu}_{X,k}^{\mathrm{T}}, \widehat{\beta}_{X,k}^{\mathrm{T}}\right)^{\mathrm{T}} = \left(\sum_{t=1}^{k} Z_{t,K} Z_{t,K}^{\mathrm{T}}\right)^{-1} \sum_{t=1}^{k} Z_{t,M} \widehat{X}_t$$

where $\widehat{X}_t = (< X_t, \widehat{\phi}_{XY}^1 >, \ldots, < X_t, \widehat{\phi}_{XY}^K >)^{\mathrm{T}}$. Analogously, we can define the recursive (least squares) estimator for θ_Y.

To compare the trends, define

$$\hat{\boldsymbol{\xi}}_k = \hat{\beta}_{X,\lfloor kN_1/N \rfloor} - \hat{\beta}_{Y,\lfloor kN_2/N \rfloor}, \quad k = 2, \ldots, N = N_1 + N_2$$

Notice that if $N_1 = N_2$, we can simply define $\hat{\boldsymbol{\xi}}_k = \hat{\beta}_{X,k} - \hat{\beta}_{Y,k}$ for $k = 2, \ldots, N_1$. We propose the following test statistic for Hypothesis IV:

$$T_N = \hat{\boldsymbol{\xi}}_N^{\mathrm{T}} \left\{\frac{1}{N^5}\sum_{k=2}^{N} k^4 \left(\hat{\boldsymbol{\xi}}_k - \hat{\boldsymbol{\xi}}_N\right)\left(\hat{\boldsymbol{\xi}}_k - \hat{\boldsymbol{\xi}}_N\right)^{\mathrm{T}}\right\}^{-1} \hat{\boldsymbol{\xi}}_N$$

It is worth noting that the normalization matrix in T_N has a different weight k^4 in comparison with the weight k^2 in the usual self-normalization statistic. It is demonstrated in our (unreported) simulation studies that such weighting scheme significantly improves the performance of the test under the null. In the following, we establish the asymptotic null distribution for T_N when $N_1/N_2 \to 1$.

Theorem 2.2 *Suppose Assumptions 6.3 and 6.4 in the Appendix hold with $\{X_t\}$ replaced by $\{\epsilon_{X,t}\}$ and $\{Y_t\}$ replaced by $\{\epsilon_{Y,t}\}$. Under Assumptions 6.6, 6.7 in the Appendix, and the $H_0 : \beta_X = \beta_Y$,*

$$T_N \to^d \tilde{\mathbf{W}}_K(1)^{\mathrm{T}} \left\{\int_0^1 r^4 \left\{\tilde{\mathbf{W}}_K(r) - \tilde{\mathbf{W}}_K(1)\right\}\left\{\tilde{\mathbf{W}}_K(r) - \tilde{\mathbf{W}}_K(1)\right\}^{\mathrm{T}} dr\right\}^{-1} \tilde{\mathbf{W}}_K(1)$$

where

$$\tilde{\mathbf{W}}_K(r) = \frac{1}{r^3}\int_0^r t\, d\mathbf{B}_K(t) - \frac{1}{2r^2}\int_0^r d\mathbf{B}_K(t)$$

Proof. See the Appendix.

Remark 2.1 Because of the linear trends in (2.1), the limiting distribution for T_N is different from those considered in Zhang and Shao (2015). We point out that the nonstandard limiting distribution in Theorem 2.2 can be approximated by

$$v_n' \left\{ \frac{1}{n^5} \sum_{k=2}^{n} k^4 (v_k - v_n)(v_k - v_n)' \right\}^{-1} v_n$$

where

$$v_k = \frac{1}{k(k^2-1)} \left(\sum_{t=1}^{k} t e_t - \frac{k+1}{2} \sum_{t=1}^{k} e_t \right)$$

for $\{e_t\} \sim^{i.i.d} N(0, \mathbf{I}_K)$, and n is a large enough positive integer, for example, 1000.

Remark 2.2 We use the following approach to obtain the preliminary estimators $\tilde{\beta}_X, \tilde{\beta}_Y, \tilde{\boldsymbol{\mu}}_X$, and $\tilde{\boldsymbol{\mu}}_Y$. Let $\{\psi_{j_1,j_2}(\mathbf{s})\}_{j_1,j_2}$ be a sequence of basis functions, such that $\psi_{j_1,j_2}(\mathbf{s}) = \tilde{\psi}_{j_1}(s_1)\tilde{\psi}_{j_2}(s_2)$ with $\{\tilde{\psi}_j(s)\}$ being the cubic b-spline basis functions. One can relabel $\{\psi_{j_1,j_2}(\mathbf{s})\}_{j_1,j_2}$ as $\{\psi_j(\mathbf{s})\}_{j=1}^{M}$. Define

$$(\tilde{\mu}_{X,1}, \ldots, \tilde{\mu}_{X,M}, \tilde{\beta}_{X,1}, \ldots, \tilde{\beta}_{X,M})^{\mathrm{T}} = \left(\sum_{t=1}^{N_1} Z_{t,M} Z_{t,M}^{\mathrm{T}} \right)^{-1} \sum_{t=1}^{N_1} Z_{t,M} \tilde{X}_t$$

where $\tilde{X}_t = (< X_t, \psi_1 >, \ldots, < X_t, \psi_M >)'$. The preliminary estimator is then given by $\tilde{\beta}_X(\mathbf{s}) = \tilde{\beta}_{X,1}\psi_1(\mathbf{s}) + \ldots + \tilde{\beta}_{X,M}\psi_M(\mathbf{s})$.

3. MONTE CARLO SIMULATIONS

We conduct simulations to evaluate the empirical sizes and powers of the proposed tests for Hypotheses III and IV. All space-time random fields in the simulation are generated from the below model,

$$X_t(s_1, s_2) = \mu_X(s_1, s_2) + \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij}(t)\phi_i(s_1)\phi_j(s_2), \quad t = 1, \ldots, T \tag{3.1}$$

where (s_1, s_2) are the two components of $\mathbf{s}$, $\mu_X(s_1, s_2)$ denotes a spatially varying mean function, $\{\phi_i(s)\}$ are cubic b-spline basis functions, and $c_{ij}(t) = \rho c_{ij}(t-1) + \epsilon_{ij}(t)$ with $\{\epsilon_{ij}(t) : i = 1, 2, \ldots, n; j = 1, \ldots, m\}$ following a zero mean Gaussian random process with covariance matrix $\boldsymbol{\Sigma}_X$. We obtain the positive definite covariance matrix $\boldsymbol{\Sigma}_X$ using an exponential covariance function $\sigma_X^2 \exp(-||\mathbf{h}||/\lambda_X)$ for $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$ for any $1 \leqslant i, j \leqslant n$. The process $Y_t(s_1, s_2)$ is also generated based on model (3.1) but with $\mu_X(s_1, s_2)$ and $\boldsymbol{\Sigma}_X$ replaced by $\mu_Y(s_1, s_2)$ and $\boldsymbol{\Sigma}_Y$ which consists of σ_Y^2 and λ_Y. The spatial locations are generated as a 21×21 grid over $[0, 1] \times [0, 1]$. We choose $m = n = 4$ and set $\rho = 0.1, 0.3, 0.6$. Each simulation setup in the succeeding text is run with $T = 250, 500, 1000$, respectively, to study the effect of sample size to the results, and all the results are based on 1000 simulations.

To evaluate the empirical sizes and powers for Hypothesis test III, we first set $\mu_X(s_1, s_2) = 0$, $\sigma_X^2 = 1$ and $\lambda_X = 0.05$. Then, we set $\mu_Y(s_1, s_2) = r_\mu Z(s_1, s_2)$ for $r_\mu = 0, 0.1, 0.2$, where $Z(s_1, s_2)$ is obtained through a Gaussian random process but was held fixed once generated. We further set $\sigma_Y^2 = 1, 1.2, 1.4$ and $\lambda_Y = 0.05, 0.1, 0.15$, respectively, to ensure that random fields Y_t can have the same/different mean and covariance structure as X_t. Different combinations of those parameter values allow us to evaluate both the sizes and powers of the test. See the details of the combinations in Tables 1–3. We choose seven eigenfunctions that correspond to 90% of total variation to report the sizes and powers. Fixing $\rho = 0.1$, Table 1 shows that the sizes are close to the nominal level 0.05, even at $T = 250$. The powers raise quickly when the mean and covariance structures of Y_t deviate from those of X_t. In order to evaluate the powers to different variances and powers to different range parameters separately, we conduct additional simulations as reported in Tables 2–3. The results indicate that either violation of the equal covariance will lead to a rejection of the test. Viewing the results for $\rho = 0.3$ and $\rho = 0.6$, we observe that when ρ increases, the sizes at a smaller number of eigenfunctions still remain correct even at $T = 250$, but at a larger number of eigenfunctions maintaining the correct sizes requires a larger T. The effect of temporal correlation on powers is also more obvious at $T = 250$. A larger ρ in general reduces the power, but a large T eliminates this effect. We omit the results for $\rho = 0.6$ in this article while presenting the results for $\rho = 0.3$ in Table 4 because the climate model data we analyze in Section 4.1 has estimated temporal autocorrelation around 0.3. In unreported simulation results, we also evaluated the powers for a special case in which μ_Y is a constant over space and takes the values 0.1,0.2, and 0.3, respectively. The results clearly show that the test is highly sensitive to this type of difference.

Table 1. Sizes and powers of testing the mean and covariance between two random fields given $\rho = 0.1$, $\boldsymbol{\mu}_X = 0$, $\sigma_X^2 = 1$ and $\lambda_X = 0.05$

$r_{\boldsymbol{\mu}}$	σ_Y^2	λ_Y	T	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7
0	1	0.05	250	0.061	0.054	0.056	0.06	0.068	0.067	0.074
0	1	0.05	500	0.042	0.040	0.039	0.053	0.054	0.047	0.055
0	1	0.05	1000	0.042	0.055	0.044	0.043	0.051	0.066	0.061
0.1	1	0.05	250	0.608	0.499	0.472	0.463	0.454	0.449	0.464
0.1	1	0.05	500	0.860	0.823	0.793	0.774	0.781	0.761	0.748
0.1	1	0.05	1000	0.972	0.963	0.961	0.975	0.985	0.984	0.979
0.2	1	0.05	250	0.960	0.964	0.973	0.983	0.986	0.983	0.989
0.2	1	0.05	500	0.997	0.997	0.999	0.999	1	1	1
0.2	1	0.05	1000	1	1	1	1	1	1	1
0	1.2	0.1	250	0.344	0.401	0.422	0.428	0.464	0.477	0.476
0	1.2	0.1	500	0.582	0.681	0.765	0.791	0.831	0.821	0.834
0	1.2	0.1	1000	0.834	0.912	0.950	0.969	0.990	0.994	0.997
0.1	1.2	0.1	250	0.733	0.721	0.731	0.748	0.763	0.778	0.780
0.1	1.2	0.1	500	0.929	0.930	0.952	0.956	0.966	0.973	0.977
0.1	1.2	0.1	1000	0.991	0.993	0.997	0.999	1	1	1
0.2	1.2	0.1	250	0.965	0.962	0.976	0.990	0.992	0.989	0.994
0.2	1.2	0.1	500	0.998	0.993	1	0.999	1	1	1
0.2	1.2	0.1	1000	1	1	1	1	1	1	1
0	1.4	0.15	250	0.873	0.939	0.978	0.981	0.994	0.991	0.993
0	1.4	0.15	500	0.983	0.994	0.998	0.998	1	1	1
0	1.4	0.15	1000	1	0.998	1	1	1	1	1
0.1	1.4	0.15	250	0.934	0.973	0.983	0.989	0.996	0.997	0.998
0.1	1.4	0.15	500	0.990	0.994	0.996	1	1	1	1
0.1	1.4	0.15	1000	1	0.999	1	1	1	1	1
0.2	1.4	0.15	250	0.992	0.989	0.998	0.996	1	0.999	1
0.2	1.4	0.15	500	0.999	1	1	1	1	1	1
0.2	1.4	0.15	1000	1	1	1	1	1	1	1

$r_{\boldsymbol{\mu}} = 0$, $\sigma_Y^2 = 1$ and $\lambda_Y = 0.05$ correspond to sizes, and all other cases correspond to powers.

Table 2. Powers of Hypothesis test III in terms of λ_Y given $\lambda_X = 0.05$

λ_Y	T	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7
0.1	250	0.075	0.088	0.085	0.091	0.096	0.079	0.087
0.1	500	0.136	0.120	0.110	0.095	0.094	0.088	0.087
0.1	1000	0.240	0.211	0.192	0.158	0.155	0.158	0.137
0.15	250	0.470	0.449	0.459	0.406	0.372	0.370	0.346
0.15	500	0.758	0.753	0.758	0.693	0.691	0.678	0.649
0.15	1000	0.932	0.925	0.965	0.96	0.969	0.963	0.945
0.2	250	0.862	0.863	0.885	0.856	0.858	0.828	0.798
0.2	500	0.982	0.982	0.988	0.987	0.984	0.986	0.991
0.2	1000	1	0.996	1	0.999	1	1	1

Table 3. Powers of Hypothesis test III in terms of σ_Y^2 given $\sigma_X^2 = 1$

σ_Y^2	T	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7
1.2	250	0.143	0.212	0.235	0.241	0.290	0.302	0.305
1.2	500	0.290	0.366	0.436	0.461	0.498	0.544	0.564
1.2	1000	0.484	0.651	0.746	0.804	0.869	0.908	0.924
1.4	250	0.441	0.561	0.659	0.710	0.804	0.838	0.861
1.4	500	0.709	0.857	0.917	0.937	0.979	0.986	0.995
1.4	1000	0.902	0.970	0.996	0.998	0.999	1	1
1.6	250	0.684	0.850	0.892	0.950	0.970	0.987	0.99
1.6	500	0.898	0.960	0.989	0.999	1	1	1
1.6	1000	0.983	0.998	1	1	1	1	1

Table 4. Sizes and powers of testing the mean and covariance between two random fields given $\rho = 0.3$ $\boldsymbol{\mu}_X = 0$, $\sigma_X^2 = 1$ and $\lambda_X = 0.05$

$r_{\boldsymbol{\mu}}$	σ_Y^2	λ_Y	T	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7
0	1	0.05	250	0.051	0.061	0.065	0.087	0.084	0.112	0.126
0	1	0.05	500	0.046	0.048	0.053	0.060	0.079	0.084	0.070
0	1	0.05	1000	0.039	0.054	0.048	0.050	0.056	0.061	0.078
0.1	1	0.05	250	0.447	0.365	0.35	0.363	0.378	0.413	0.440
0.1	1	0.05	500	0.722	0.662	0.618	0.608	0.621	0.625	0.621
0.1	1	0.05	1000	0.919	0.894	0.893	0.893	0.914	0.900	0.900
0.2	1	0.05	250	0.897	0.878	0.918	0.921	0.932	0.948	0.958
0.2	1	0.05	500	0.987	0.984	0.988	0.991	0.996	0.997	0.999
0.2	1	0.05	1000	0.999	1	1	1	0.999	1	1
0	1.2	0.1	250	0.319	0.363	0.398	0.422	0.470	0.507	0.533
0	1.2	0.1	500	0.524	0.617	0.694	0.728	0.791	0.794	0.806
0	1.2	0.1	1000	0.783	0.853	0.925	0.954	0.976	0.990	0.991
0.1	1.2	0.1	250	0.595	0.608	0.628	0.664	0.702	0.722	0.762
0.1	1.2	0.1	500	0.864	0.875	0.910	0.909	0.937	0.945	0.951
0.1	1.2	0.1	1000	0.968	0.980	0.988	0.997	0.998	1	1
0.2	1.2	0.1	250	0.923	0.917	0.929	0.955	0.961	0.976	0.979
0.2	1.2	0.1	500	0.988	0.989	0.997	0.996	0.998	0.999	0.999
0.2	1.2	0.1	1000	1	1	1	1	1	1	1
0	1.4	0.15	250	0.841	0.919	0.959	0.974	0.986	0.992	0.993
0	1.4	0.15	500	0.972	0.985	0.995	0.999	1	1	1
0	1.4	0.15	1000	0.996	0.999	0.999	1	1	1	1
0.1	1.4	0.15	250	0.909	0.945	0.971	0.983	0.993	0.997	0.998
0.1	1.4	0.15	500	0.980	0.992	0.998	0.999	1	1	1
0.1	1.4	0.15	1000	0.998	0.999	1	1	1	1	1
0.2	1.4	0.15	250	0.971	0.976	0.990	0.997	1	0.999	1
0.2	1.4	0.15	500	0.997	0.997	1	1	1	1	1
0.2	1.4	0.15	1000	1	1	1	1	1	1	1

$r_{\boldsymbol{\mu}} = 0$, $\sigma_Y^2 = 1$ and $\lambda_Y = 0.05$ correspond to sizes, and all other cases correspond to powers.

Table 5. Sizes and powers of testing the trend between two random fields

ρ	r_β	T	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7
0.1	1	250	0.030	0.045	0.043	0.054	0.053	0.051	0.062
	1	500	0.069	0.065	0.063	0.059	0.064	0.066	0.078
	1	1000	0.057	0.051	0.049	0.035	0.044	0.044	0.066
	1.05	250	0.525	0.435	0.391	0.357	0.337	0.327	0.364
	1.05	500	0.998	0.994	0.989	0.986	0.989	0.989	0.993
	1.05	1000	1	1	1	1	1	1	1
	1.1	250	0.959	0.914	0.904	0.897	0.883	0.884	0.925
	1.1	500	1	1	1	1	1	1	1
	1.1	1000	1	1	1	1	1	1	1
0.3	1	250	0.030	0.051	0.046	0.06	0.063	0.066	0.080
	1	500	0.072	0.07	0.069	0.063	0.072	0.073	0.084
	1	1000	0.058	0.055	0.049	0.037	0.041	0.046	0.070
	1.05	250	0.379	0.316	0.287	0.26	0.255	0.244	0.302
	1.05	500	0.989	0.961	0.953	0.956	0.949	0.945	0.965
	1.05	1000	1	1	1	1	1	1	1
	1.1	250	0.871	0.801	0.777	0.769	0.747	0.759	0.807
	1.1	500	1	1	1	1	1	1	1
	1.1	1000	1	1	1	1	1	1	1
0.6	1	250	0.042	0.064	0.066	0.087	0.106	0.105	0.141
	1	500	0.084	0.087	0.082	0.076	0.091	0.096	0.104
	1	1000	0.058	0.057	0.056	0.044	0.052	0.055	0.084
	1.05	250	0.22	0.205	0.172	0.188	0.198	0.197	0.252
	1.05	500	0.862	0.802	0.791	0.764	0.766	0.764	0.789
	1.05	1000	1	1	1	1	1	1	1
	1.1	250	0.599	0.548	0.504	0.49	0.496	0.509	0.564
	1.1	500	0.999	1	0.999	0.993	0.994	0.994	0.997
	1.1	1000	1	1	1	1	1	1	1

$r_\beta = 1$ corresponds to sizes; $r_\beta > 1$ corresponds to powers.

To evaluate the sizes and powers for Hypothesis test IV, we set $E\{X_t(s_1, s_2)\} = \beta_X(s_1, s_2)t$ with $\beta_X(s_1, s_2)$ generated from the following model:

$$\beta_X(s_1, s_2) = \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}\phi_i(s_1)\phi_j(s_2)$$

where $\{c_{ij} : i = 1, 2, \ldots, n; j = 1, 2, \ldots, m\}$ is generated from a Gaussian random process with an exponential covariance function but is held fixed throughout the simulation once obtained. The generated $\beta_X(s_1, s_2)$ is scaled to reflect the magnitude of actual trend of the observed climate. We then set $E\{Y_t(s_1, s_2)\} = \beta_Y(s_1, s_2)t$ with $\beta_Y(s_1, s_2) = r_\beta\beta_X(s_1, s_2)$. By setting $r_\beta = 1, 1.05, 1.1$, respectively, we are allowed to evaluate both the sizes and powers of comparing the trend between two random fields. Table 5 reports the simulation results for different combinations of r_β and T at $\rho = 0.1, 0.3, 0.6$, respectively. It is seen that when ρ is moderately low, the sizes are close to the nominal level 0.05 even at $T = 250$ and the powers of the test are satisfactory even at small differences such as $r = 1.1$. When ρ is large, the sizes at a small number of eigenfunction still remain correct, yet sizes at a large number of eigenfunctions are elevated for small T. In the latter case, a large T is helpful to retain the appropriate sizes. The powers at large ρ also become weakened compared with those at small ρ if T is small, but the powers increase quickly as T increases. It is worth mentioning that the first-order temporal autocorrelation for reanalysis data and climate modeled data in Section 4.2 is between 0.3 and 0.4.

4. APPLICATION TO SYNTHETIC CLIMATE

We consider last-millennium and historical simulations from five climate modeling centers as configured and implemented in the coupled model intercomparison project Phase 5 and the paleoclimate modelling intercomparison project Phase 3 (CMIP5/PMIP3): the Beijing Climate Center CSM1.1 model (hereinafter BCC), the National Center for Atmospheric Research Community Climate System version 4 model

(hereinafter CCSM), the Goddard Institute for Space Studies E2-R model (hereinafter GISS), the Institute Pierre-Simon Laplace CM5A-LR model (hereinafter IPSL), and the Max Plank Institute ESM-LR model (hereinafter MPI). The last-millennium simulations span the period 850–1850 CE and are forced with reconstructed time-varying radiative forcings (Schmidt *et al.*, 2011). We also use the first ensemble member of the CMIP5 historical runs (1850–2005 CE) from each of the five GCMs that were used to perform the last millennium simulations earlier. The annual means of the modeled surface temperature fields are all interpolated to even 5-degree latitude–longitude grids.

We first assess whether climate models generate statistically equivalent climates and then investigate the coherency between modeled climate and observationally based data. All testing results are subject to the number of leading eigenfunctions, also called empirical orthogonal functions (EOFs) in climatology, or principal components (PCs) that are chosen prior to the test. The testing results therefore can be interpreted as being associated with the percentage of total variation that are explained by the selected PCs. Because the percentage of total variation is a more informative reflection of the data capacity than the number of leading PCs, we present the p-values of the tests along with the percentage of total variation.

Following the tradition of functional data analysis, we smooth each of the datasets aforementioned before applying the test. The smoothing procedure ensures that our method is applicable regardless of whether the two random fields are observed at the same set of locations. Specifically, suppose our raw data are discrete observations $\{X_t(\mathbf{s}_j)\}_{j=1}^n$ and $\{Y_t(\mathbf{s}'_j)\}_{j=1}^m$ for $\mathbf{s}_j, \mathbf{s}'_j \in D$. Using smoothing spline, we first translate these discrete observations into smooth surfaces/images over the space D. We then perform our analysis on the smooth functional objects. Therefore, our tests allow the two random fields in comparison to be observed at different locations, that is, $\{\mathbf{s}_j\}_{j=1}^n$ and $\{\mathbf{s}'_j\}_{j=1}^m$ can be different. The smoothness of our data is obtained through a set of 120 basis functions that are formed by 12 cubic b-splines in the longitude and 10 cubic b-splines in the latitude direction.

4.1. Comparison between climate models

In order to focus on testing the similarity of internal variability between climate models, we only use the last-millennium simulations (850–1850 CE) in the comparison. This is because on the one hand the greenhouse gas forcings impose dominating influence on the model simulations after 1850, and on the other hand, we have reserved the assessment of trend for the posterior 1850 period in the next section. There are $\binom{5}{2} = 10$ different pairs of modeled climate fields. For each pair of models, we first remove their common annual temperature average to attain the stationarity in time and then perform $TS1$, $TS2$, and $TS3$ at a sequential number of PCs, ranging from one to the minimal number that reaches 85% of the cumulative variance ratio. The average number of required PCs to reach 85% across the 10 pairs is about 38.

We use EOFs as the basis functions in our hypothesis test. As an example of EOF-PC decompositions and comparisons among models, Figure 1 plots the first three EOFs for the GISS and IPSL models, their joint EOFs which are computed from the pooled sample covariance matrix and the respective PCs for the model EOFs and their joint EOFs. The EOF patterns are characteristic of the leading patterns of variability in global temperature fields, which are the result of different semi-oscillatory modes including the El Niño Southern Oscillation, the Arctic Oscillation, the Pacific Decadal Oscillation, and large-scale mean variability. It is important to note that EOFs reflecting these patterns of variability are ordered differently in each model, which reflects the differences in the simulated dynamics within each model. The large-scale mean variability is expressed largely as a singular leading pattern in the GISS simulation, whereas the pattern is mixed across the leading patterns in the IPSL simulation. This can be a characteristic of the model sensitivities to large-scale radiative forcings and the magnitude of internal variability associated with different simulated atmosphere-ocean phenomena in the models. For our purposes herein, it

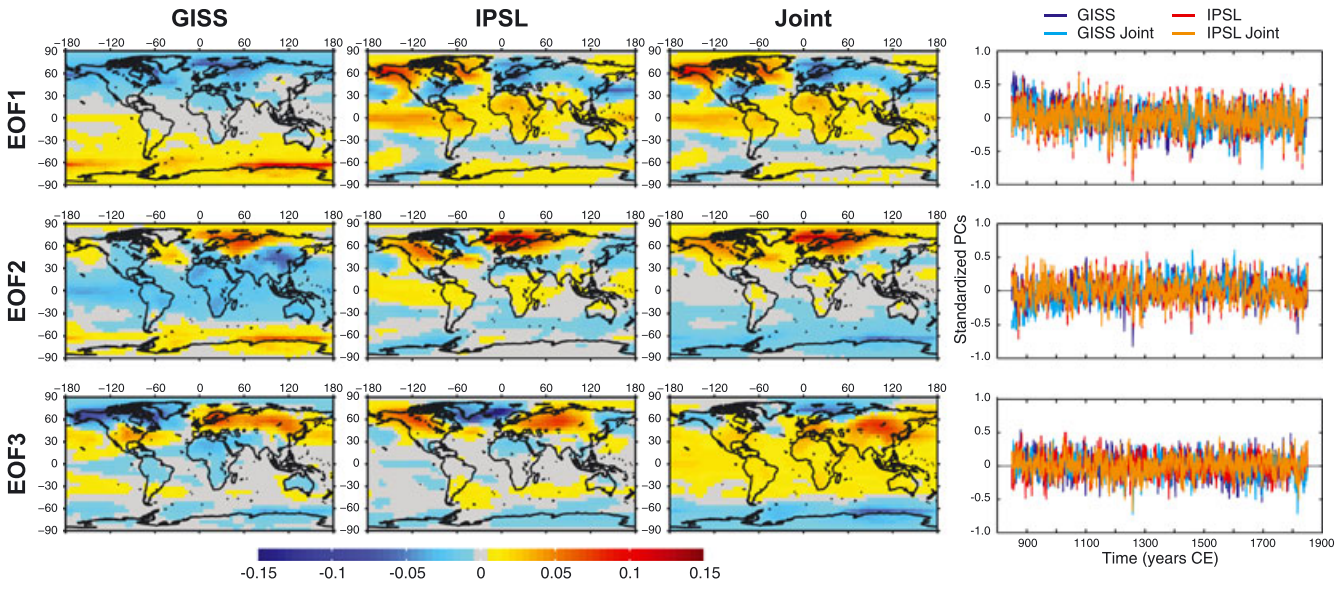

Figure 1. The first three eigenfunctions (or EOFs) and principal components of Goddard Institute for Space Studies (GISS), Institute Pierre–Simon Laplace (IPSL), and their joint field

is sufficient to note that each individual basis function in our test is the product of a certain combination of dynamics and forcing sensitivities in each model. Our motivation herein, however, is to determine the degree to which model fields are the same. Comparing only equally ranked EOFs and PCs is therefore appropriate for this evaluation.

To summarize the ensemble of temperature field comparisons across all models, Figure 2 reports the p-values of hypothesis tests, $TS1$ and $TS2$, between all simulated temperature fields along a sequence of total variation percentages. The p-values of hypothesis test, $TS3$, are all nearly zero, and thus, the plot for $TS3$ is omitted. All pairwise mean comparisons yield small p-values. This indicates that the mean surface of climate model simulations are all different, even when projected onto any subspace defined by their joint eigenfunctions. However, in terms of covariance comparison, five out of 10 pairs have similar covariance structure at the direction of their first joint eigenfunction, and among the five pairs the BCC-IPSL and BCC-MPI continue to show similarity up to their first two joint eigenfunctions, which correspond to roughly 19% of total variation in the data. The IPSL-MPI pair also shows somewhat similar covariance structure at their first two eigenfunctions. Unsurprisingly, the comprehensive test $TS3$ indicates that no single pair of modeled climate shares the common first and second moments.

4.2. Assessment of climate models using reanalysis data

An important question in climate model evaluation is to examine the performance of climate models relative to the observed climate. We address this question by comparing the simulations of the five aforementioned CMIP5 models with the surface temperature field derived from the 20th-century reanalysis product, which assimilates observations of synoptic pressure, sea surface temperatures, and sea ice from 1871 to 2012 (Compo *et al.*, 2011). The reanalysis data are presented over an even two-degree latitude–longitude grid rather than the five-degree grid of the model simulations. But because we only compare the smoothed data based on b-spline bases, the actual observations of two random fields are unimportant. The most striking feature of the temperature field over the reanalysis interval is its upward mean trend into the 21st century. We therefore perform a trend assessment during the overlapping period (1871–2005) between the reanalysis data and the model simulations as an evaluation of whether the models respond to greenhouse gas forcings in similar fashions.

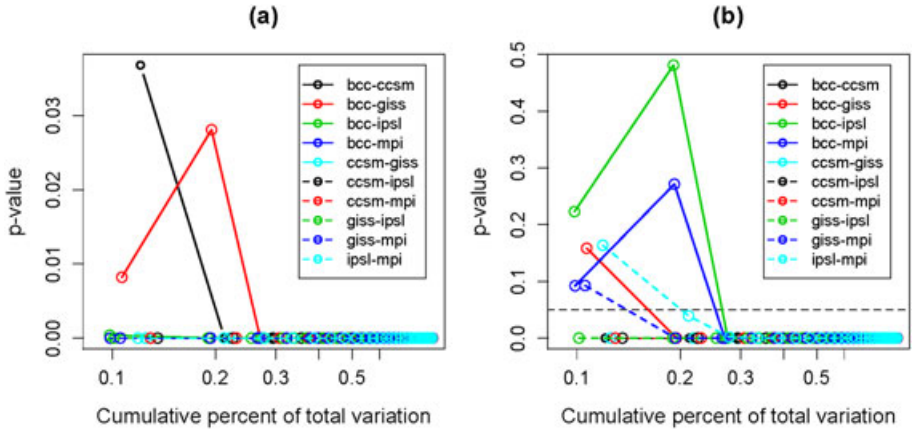

Figure 2. *P*-values from comparisons of modeled climate: (a) test for mean surface; (b) test for covariance function. The grey dashed line represents p-value equals 0.05, and the horizontal axis is in logarithmic scale

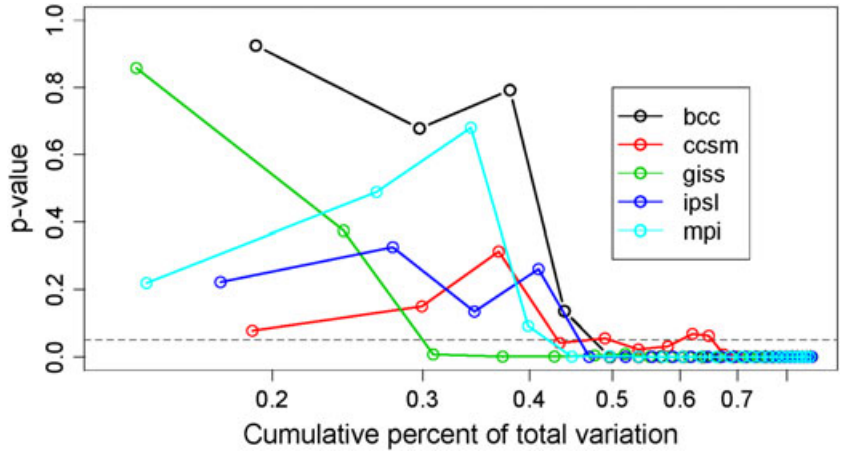

Figure 3. *P*-values from comparisons of the trend between climate models and reanalysis data. The grey dashed line represents p-value equals 0.05, and the horizontal axis is in logarithmic scale

We conduct the hypothesis test described in Section 2.2 to compare the trend in each of the five climate models to that of the reanalysis data. Figure 3 shows the p-values for the trend assessments. All five models seem to follow the trend of reanalysis data to some extent, but when projected to a large subspace that contains 70% of total variation of the data, then there is evidence that the trend of each climate model deviates from that of the reanalysis data. Despite some of its p-values being slightly below 0.05, the CCSM simulation appears to outperform the other four climate models by retaining the trend of reanalysis data to the largest capacity, followed by BCC, IPSL, MPI, and GISS in a descending order.

Note that all of the models and observational data are subject to the phasing of internal climate variability, or the timing of natural oscillations like El Niño Southern Oscillation, the Pacific Decadal Oscillation, and the Atlantic-multidecadal oscillation. Because these phenomena have variability on decadal and longer timescales that influence global temperatures, they can affect estimates of 20th-century temperature trends. While we do not account for those influences in the tests that we have performed, a more robust assessment across all of the ensemble members of a given CMIP5 model simulation would allow our test to more fully characterize the range of agreement for an individual model and the associated impacts on cross-model assessments.

5. SUMMARY AND DISCUSSION

The comparison methods that we have developed and applied herein account for the specific properties of observed and modeled climate, thus making them ideal tools for the problems of interest. Alternative tests are of course additionally applicable, such as comparisons based on the squared differences of the mean or covariance functions from the two random fields (Benko *et al.*, 2009). These alternative approaches can be viewed as a simultaneous test that integrates the differences in all different directions with equal weights. For our data sets, it might be expected that the null hypothesis will be rejected in all cases using this test, based on the behavior of p-values at large percentages of total variation in our results. Such a global test therefore fails to provide details of the comparison between two random fields, such as the source or direction of the differences. Another disadvantage of a test based on squared differences is that the limiting distribution is non-pivotal as it depends on the second or fourth-order structures of the functional time series and therefore it has to resort to the bootstrap or subsampling calibration (Benko *et al*., 2009). Finally, the tests based on mean squared differences require the choice of tuning parameters, making our tests more informative and convenient to use. The advances in the current paper are also more rigorous than those developed by Li and Smerdon (2012), although both have focused on the first and second-moment structures of the investigated fields. Compared with Li and Smerdon (2012), the methods developed herein have relaxed several unrealistic assumptions, such as the spatial stationarity and temporal independence. Moreover, our methods are nonparametric and thus avoid the risk of model misspecification.

We have provided an example assessment for five climate model simulations spanning the 850–2005 CE interval. The assessment includes the characterization of differences between climate model simulations, and the skill evaluation of model simulations in reproducing the main features of observationally based reanalysis data. Because the modeled climates are high-dimensional data ($p > n$), it is necessary to reduce the dimension in order to focus on the primary characteristics of the climate field. Our evaluations and comparisons were thus conducted in low-dimensional spaces using the functional data analysis approach. The projections of the data onto those low dimensions retain the majority of the variability in the original data, and more importantly, the directions of those projections can correspond to certain scientific interpretations. Among the simulated temperature fields considered herein, none of them are fully equivalent in terms of their underlying first and second moment structures. Some show equivalences, however, of the covariance in reduced dimensional space. In terms of capturing the upward trend of the 20th century, the CCSM mostly matches with the reanalysis data. Our conclusions are based on evaluating the differences between the mean surfaces, the covariance structures, and the trend surfaces rather than evaluating only the global mean differences of the surfaces, the latter of which is often seen in the climate literature. Our findings are indicative of the fact that different climate models in general have different climate sensitivities and internal dynamics, thus generating climates of different characteristics; but some can be similar in terms of the major components of their dynamics and climate sensitivity.

Acknowledgements

Li's research was partially supported by NSF-DPP-1418339, and Smerdon's research is partially supported by Lamont contribution #8009. The authors thank the editor, and the referees for constructive suggestions that have improved the content and presentation of this article.

REFERENCES

Benko M, Härdle W, Kneip A. 2009. Common functional principal components. *Annals of Statistics* **37**:1–34.
Bosq D. 2000. *Linear Process in Function Spaces: Theory and Applications*. Springer: New York.
Briggs WM, Levine RA. 1997. Wavelets and field forecast verification. *Monthly Weather Review* **125**:1329–1341.
Compo GP, Whitaker JS, Sardeshmukh PD, Matsui N, Allan RJ, Yin X, Gleason BE, Vose RS, Rutledge G, Bessemoulin P, Brnnimann S, Brunet M, Crouthamel RI, Grant AN, Groisman PY, Jones PD, Kruk M, Kruger AC, Marshall GJ, Maugeri M, Mok HY, Nordli Ø, Ross TF, Trigo RM, Wang XL, Woodruff SD, Worley SJ. 2011. The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society* **137**:1–28.
Cuevas A, Febrero M, Fraiman R. 2004. An ANOVA test for functional data. *Computational Statistics & Data Analysis* **47**:111–122.
Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**:253–263.
Fan J, Lin S-K. 1998. Tests of significance when the data are curves. *Journal of the American Statistical Association* **93**:1007–1021.
Fremdt S, Horváth L, Kokoszka P, Steinebach JG. 2013. Testing the equality of covariance operators in functional samples. *Scandinavian Journal of Statistics* **40**:138–152.
Hering A, Genton MG. 2011. Comparing spatial predictions. *Technometrics* **53**:414–425.
Hormann S, Kokoszka P. 2010. Weakly dependent functional data. *The Annals of Statistics* **38**:1845–1884.
Horvath L, Kokoszka P. 2012. *Inference for Functional Data with Applications*. Springer: New York.

Horváth L, Kokoszka P, 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.
Kraus D, Panaretos VM. 2012. Dispersion operators and resistant second-order functional data analysis. *Biometrika* **99**:813–832.
Li B, Smerdon J. 2012. Defining spatial comparison metrics for evaluation of paleoclimatic field reconstructions of the common Era. *Environmetrics* **23**:394–406.
Lund R, Li B. 2009. Revisiting climate region definitions via clustering. *Journal of Climate* **22**:1787–1800.
Panaretos VM, Kraus D, Maddocks JH. 2010. Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *Journal of the American Statistical Association* **105**:670–682.
Pavlicová M, Santer TJ, Cressie N. 2008. Detecting signals in FMRI data using powerful FDR procedures. *Statistics and Its Interface* **1**:23–32.
Schmidt GA, Jungclaus JH, Ammann CM, Bard E, Braconnot P, Crowley TJ, Delaygue G, Joos F, Krivova NA, Muscheler R, Otto-Bliesner BL, Pongratz J, Shindell DT, Solanki SK, Steinhilber F, Vieira LEA. 2011. Climate forcing reconstructions for use in PMIP simulations of the last millennium (v1.0). *Geoscientific Model Development* **4**:33–45. DOI: 10.5194/gmd-4-33-2011.
Shen X, Huang HC, Cressie N. 2002. Nonparametric hypothesis testing for a spatial signal. *Journal of the American Statistical Association* **97**:1122–1140.
Snell SE, Gopal S, Kaufmann RK. 2000. Spatial interpolation of surface air temperatures using artificial neural networks: evaluating their use for downscaling GCMs. *Journal of Climate* **13**:886–895.
Wang W, Anderson BT, Entekhabi D, Huang D, Su Y, Kaufmann RK, Potter C, Myneni RB. 2007. Intraseasonal interactions between temperature and vegetation over the boreal forests. *Earth Interactions* **11**:1–30.
Zhang X, Shao X. 2015. Two sample inference for the second-order property of temporally dependent functional data. *Bernoulli* **21**:909–929.

APPENDIX

Definition 6.1 Assume that $\{U_i\} \in L^p_{\mathbb{H}}$ with $p > 0$ admits the following representation:

$$U_i = f(\varepsilon_i, \varepsilon_{i-1}, \dots), \ i = 1, 2, \dots, \tag{A1}$$

where the ε_i's are iid elements taking values in a measurable space S and f is a measurable function $f : S^\infty \to \mathbb{H}$. For each $i \in \mathbb{N}$, let $\{\varepsilon_j^{(i)}\}_{j\in\mathbb{Z}}$ be an independent copy of $\{\varepsilon_j\}_{j\in\mathbb{Z}}$. The sequence $\{U_i\}$ is said to be L^p-m-approximable if

$$\sum_{m=1}^{\infty} (E||U_m - U_m^{(m)}||^p)^{1/p} < \infty \tag{A2}$$

where $U_i^{(m)} = f(\varepsilon_i, \varepsilon_{i-1}, \dots, \varepsilon_{i-m+1}, \varepsilon_{i-m}^{(i)}, \varepsilon_{i-m-1}^{(i)}, \dots)$.

Assumption 6.1 Assume $N_1/N \to \gamma_1$ and $N_2/N \to \gamma_2$ as $\min(N_1, N_2) \to +\infty$, where $\gamma_1, \gamma_2 \in (0, 1)$.

Assumption 6.2 Assume $\{X_t(\mathbf{s})\}_{t=1}^{+\infty} \subseteq L^2_{\mathbb{H}}$ and $\{Y_t(\mathbf{s})\}_{t=1}^{+\infty} \subseteq L^2_{\mathbb{H}}$ are both L^4-m-approximable, and they are mutually independent.

Assumption 6.3 Assume $\{(X_t(\mathbf{s}), Y_t(\mathbf{s}))\}_{t=1}^{+\infty} \subseteq L^4_{\mathbb{H}\times\mathbb{H}}$ is an L^4-m-approximable sequence.

Assumption 6.4 Let $\{\lambda_X^j\}$ and $\{\lambda_Y^j\}$ be the eigenvalues associated with C_X and C_Y, respectively. Assume $\lambda_X^1 > \lambda_X^2 > \ldots > \lambda_X^{K+1}$ and $\lambda_Y^1 > \lambda_Y^2 > \cdots > \lambda_Y^{K+1}$, for some positive integer $K \geqslant 2$.

Denoted by $\{\tilde{\phi}_{XY}^i\}$, the eigenfunctions of $\gamma_1 C_X + \gamma_2 C_Y$. Let $V_{X,t} = (< X_t - \boldsymbol{\mu}_X, \tilde{\phi}_{XY}^1 >, \dots, < X_t - \boldsymbol{\mu}_X, \tilde{\phi}_{XY}^K >)^{\mathrm{T}}$ and $R_{X,t} = (< \{(X_t - \boldsymbol{\mu}_X) \otimes (X_t - \boldsymbol{\mu}_X) - C_X\}\tilde{\phi}_{XY}^i, \tilde{\phi}_{XY}^j >)_{1\leqslant i,j\leqslant K}$. Define $W_{X,t} = (V_{X,t}^{\mathrm{T}}, \mathrm{vech}(R_{X,t})^{\mathrm{T}})^{\mathrm{T}}$ and the analogous quantity $W_{Y,t}$ for the second sample, where vech is the operator that stacks the elements on and below the main diagonal of a symmetric $K \times K$ matrix as a vector with $K(K+1)/2$ components.

Assumption 6.5 Assume the asymptotic covariance matrix of

$$\frac{1}{\sqrt{N}}\left(\frac{1}{\gamma_1}\sum_{t=1}^{N_1} W_{X,t} - \frac{1}{\gamma_2}\sum_{t=1}^{N_2} W_{Y,t}\right)$$

is positive definite.

Assumption 6.6 Assume the asymptotic covariance matrix of

$$\frac{1}{\sqrt{N}}\left(\sum_{t=1}^{N_1} \breve{\epsilon}_{X,t} - \sum_{t=1}^{N_2} \breve{\epsilon}_{Y,t}\right)$$

is positive definite.

Assumption 6.7 Assume that for $1 \leqslant j \leqslant K$, $\limsup_{N\to+\infty} N||\phi_{XY}^j - \widehat{C}_j \widehat{\phi}_{XY}^j||^2 < \infty$, where $\widehat{C}_j = \mathrm{sign}(< \widehat{\phi}_{XY}^j, \phi_{XY}^j >)$.

The basic idea in Definition 6.1 is to approximate a stationary sequence with random variables that exhibit finite dependence. The L^p-m-approximable condition in Assumptions 6.2–6.3 is satisfied for many functional time series models such as functional autoregressive models, functional bilinear models, and functional autoregressive conditional heteroskedasticity models, see more details in (Hormann and Kokoszka, 2010). Assumption 6.4 assumes distinct eigenvalues, which are common in the literature, see for example, (Bosq, 2000) and (Horvath and Kokoszka, 2012). Assumptions 6.5–6.6 assume the asymptotic covariance matrices to be positive definite, which is relatively mild. Finally, we also require the estimated eigenfunctions to be $\sqrt{n}$-consistent in Assumption 6.7.

Proof of Theorem 2.2. Denote by $N_0 = N_1 = N_2$. Let $\mathbf{D} = \text{diag}(N_0^{-1/2}\mathbf{1}_K, N_0^{-3/2}\mathbf{1}_K)$ where $\mathbf{1}_K = (1, 1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^K$. Because $\{\epsilon_{X,t}\}$ is a L^4-m-approximable sequence, $\{\check{\epsilon}_{X,t}\}$ is a L^4-m-approximable sequence as well (Hörmann and Kokoszka, 2010). Using summation and integration by parts, and the continuous mapping theorem, we have for $k = \lfloor N_0 r \rfloor$,

$$\left(\sum_{t=1}^{k} \mathbf{D} Z_t Z_t^{\mathrm{T}} \mathbf{D}\right)^{-1} \to \begin{pmatrix} r\mathbf{I}_K & (r^2/2)\mathbf{I}_K \\ (r^2/2)\mathbf{I}_K & (r^3/3)\mathbf{I}_K \end{pmatrix}^{-1} = \begin{pmatrix} (4/r)\mathbf{I}_K & (-6/r^2)\mathbf{I}_K \\ (-6/r^2)\mathbf{I}_K & (12/r^3)\mathbf{I}_K \end{pmatrix}$$

and

$$\mathbf{D}\sum_{t=1}^{k} Z_t \check{\epsilon}_{X,t} \Rightarrow^d \begin{pmatrix} \Lambda_j \int_0^r d\mathbf{B}_K(t) \\ \Lambda_j \int_0^r t d\mathbf{B}_K(t) \end{pmatrix}$$

where $\mathbf{B}_K(t)$ denotes the K-dimensional vector of independent standard Brownian motions, Λ_j is the matrix square root of the long run variance matrix of $\check{\epsilon}_{X,t}$, and " $\Rightarrow^d$ " denotes weak convergence in a functional space. Thus, we have

$$\begin{aligned} \mathbf{D}^{-1}(\check{\theta}_{X,k} - \theta_X) &= \left(\sum_{t=1}^{k} \mathbf{D} Z_t Z_t^{\mathrm{T}} \mathbf{D}\right)^{-1} \mathbf{D} \sum_{t=1}^{k} Z_t \check{\epsilon}_{X,t} \\ &\Rightarrow^d \begin{pmatrix} (4/r)\mathbf{I}_K & (-6/r^2)\mathbf{I}_K \\ (-6/r^2)\mathbf{I}_K & (12/r^3)\mathbf{I}_K \end{pmatrix} \begin{pmatrix} \Lambda_j \int_0^r d\mathbf{B}_K(t) \\ \Lambda_j \int_0^r t d\mathbf{B}_K(t) \end{pmatrix} \end{aligned}$$

It implies that

$$N_0^{3/2}(\check{\beta}_{X,k} - \check{\beta}_X) \Rightarrow^d 12\Lambda_j \left(\frac{1}{r^3}\int_0^r t d\mathbf{B}_K(t) - \frac{1}{2r^2}\int_0^r d\mathbf{B}_K(t)\right) = 12\Lambda_j \tilde{\mathbf{W}}_K(r)$$

where

$$\tilde{\mathbf{W}}_K(r) = \frac{1}{r^3}\int_0^r t d\mathbf{B}_K(t) - \frac{1}{2r^2}\int_0^r d\mathbf{B}_K(t)$$

The same argument applies to the second sample. Define $\check{\mathbf{m}}\zeta_k = \check{\beta}_{X,k} - \check{\beta}_{Y,k}$ for $k = 2, \ldots, N_0$. Under $H_0 : \beta_X = \beta_Y$, we have

$$N_0^{3/2}\check{\mathbf{m}}\zeta_k \Rightarrow 12\tilde{\Lambda}\tilde{\mathbf{W}}_K(r)$$

where $\tilde{\Lambda}$ is the matrix square root of the long run variance matrix of $\check{\epsilon}_{X,t} - \check{\epsilon}_{Y,t}$. By the continuous mapping theorem, we obtain

$$\begin{aligned} \check{T}_N &= \check{\mathbf{m}}\zeta_{N_0}^{\mathrm{T}} \left\{\frac{1}{N_0^5}\sum_{k=2}^{N_0} k^4 (\check{\mathbf{m}}\zeta_k - \check{\mathbf{m}}\zeta_{N_0})(\check{\mathbf{m}}\zeta_k - \check{\mathbf{m}}\zeta_{N_0})^{\mathrm{T}}\right\}^{-1} \check{\mathbf{m}}\zeta_{N_0} \\ &\to^d \tilde{\mathbf{W}}_K(1)^{\mathrm{T}} \left\{\int_0^1 r^4 \{\tilde{\mathbf{W}}_K(r) - \tilde{\mathbf{W}}_K(1)\}\{\tilde{\mathbf{W}}_K(r) - \tilde{\mathbf{W}}_K(1)\}^{\mathrm{T}} dr\right\}^{-1} \tilde{\mathbf{W}}_K(1) \end{aligned} \tag{A3}$$

Under Assumption 6.7, we can show that the estimation effect by replacing ϕ_{XY}^j with $\widehat{\phi}_{XY}^j$ is asymptotically negligible (see, for example, the proof of Theorem A.2 of Hörmann and Kokoszka 2010). Thus, T_N converges to the same limit in (A3).

