Environmental and Ecological Statistics manuscript No.
(will be inserted by the editor)

Modeling Regional Impacts of Climate Teleconnections using Functional Data Analysis

Simon J Bonner, Nathaniel K Newlands, and Nancy E Heckman

Received: date / Accepted: date

Abstract Teleconnections are quasi-periodic changes in atmospheric circulation that oscillate over long periods of time and impact climate over large regions. These patterns are often linked to long-term variations in climate and extreme weather events and may explain regional differences in climate vulnerability. We apply methods of functional data analysis (FDA) to examine regional impacts of teleconnections on climate in British Columbia, Canada, between 1951 and 2000. We focus on monthly mean temperature as an overall determinant of crop growth and apply functional principal components analysis (FPCA) to study variations in the impacts of four major teleconnection indices affecting the Northern Hemisphere (the Southern Oscillation Index (SOI), the Pacific North American (PNA), Pacific Decadal Oscillation (PDA), and the North American Oscillation (NAO) indices). Two challenges we consider are that the impacts of teleconnections cannot be observed directly and that fine scale data required to study regional variations may come from different sources with highly varied records. We first fit thin-plate regression splines to the raw data to construct complete series of pseudo-data at fixed grid points. Regression models incorporating Bayesian P-splines were then fit to the pseudo-data to estimate the impacts of the four teleconnections over time. Finally, FPCA was then applied to study regional variations in these effects. Our analysis identified strong variations in mean temperature associated with the PNA. The resulting spatial patterns also reveal areas of

S.J. Bonner
Department of Statistics, University of Kentucky, Lexington, KY, 40536-0082, USA
Tel.: +1-859-257-4950
Fax: +1-859-323-1973
E-mail: simon.bonner@uky.edu

N.K. Newlands
Science and Technology Branch, Agriculture and Agri-Food Canada, Lethbridge Research Centre, PO Box 3000, Lethbridge, AB, T1J 4B1, Canada

N.E. Heckman
Department of Statistics, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada

increased/decreased temperature variability that may have higher climate risk or be suitable for expansion of agricultural activity.

Keywords Agriculture; Climate Variability; Functional Data Analysis; Functional Principal Components Analysis; Teleconnections

1 Introduction

Teleconnections are large scale, quasi-periodic patterns in atmospheric circulation induced by changes in air pressure and sea surface temperature that affect climate over large regions of the earth. Although their oscillations are not strictly periodic, these phenomena have been associated with repeated fluctuations in temperature and rainfall across the globe. The best known teleconnection which affects North American climate is the El-Ñino Southern Oscillation (ENSO). Measured by differences in air pressure between Tahiti and Darwin, Australia, ENSO is associated with broad changes in atmospheric pressure that alter the movement of warm and cold air masses over the Pacific Ocean. Oscillations occur every 3 to 7 years and the extreme phases, known as El-Ñino and La-Ñina, have well known effects on climate across the Americas and along the Western Pacific. In this analysis, we apply methods of functional data analysis (FDA) to examine associations between climate in British Columbia (BC), Canada, and indices of four teleconnection patterns impacting western Canada to better understand their regional effects and their potential impacts on agriculture. The indices we consider are the Southern Oscillation Index (SOI), the Pacific North-American pattern (PNA), the North-Atlantic Oscillation (NAO), and the Pacific-Decadal Oscillation (PDO) (Wang et al, 2006; Sheridan, 2003; Shabbar and Khandekar, 1996).

Understanding the effects of teleconnections on local climate is critical to predicting crop yields and maintaining reliable food supplies. Teleconnections explain large portions of the variation in annual air circulation patterns. They persist over large enough areas and long enough time scales that they may alter survival thresholds in the phenological response of crops. Teleconnections also impact air temperature, precipitation, and other agro-climate indices that affect seasonal crop growth including first fall frost day, growing degree day, crop heat units, and root-zone soil moisture. Models of teleconnections and their impacts on climate may also be combined with remote-sensing data (e.g., the net-difference vegetation index (NDVI)) to improve forecasts of in-season crop yield and long-term production over the next 10, 20, or even 50 years. Such models will help to prevent supply risks and to predict extreme climatic events (floods and droughts) that can affect major regions of crop production. Recent evidence has shown that teleconnections can also explain spatial patterns in the outbreak of diseases and pests that impact food production (e.g., Kriss et al, 2012). Improving the abilities of Canadian agriculture managers to reduce crop risks and ensure a sustainable food supply is critical given that Canada is a major exporter of a wide range of agricultural products including spring wheat, coarse grains (e.g., barley and corn), oilseeds (e.g., canola and flaxseed), specialty crops (e.g., lentils and chickpeas), and other horticultural products in BC.

Functional data analysis has been described as the extension of multivariate methods to infinite dimensional spaces of functions. Data appropriate for a multivariate analysis comprise finite length vectors observed from a set of subjects. The elements of these vectors may have different scales and need not relate to the same process, but the vectors must be comparable (i.e., they must be the same length and the elements must correspond across the subjects). Multivariate methods do not provide any advantage if the elements are measurements of the same process and cannot account for functional structures like smoothness. In comparison, data appropriate for FDA comprise vectors of observations of a single process indexed by a second variable – often time. Each vector is considered as a surrogate for a latent, individual specific function which is the real object of interest. It is not necessary for individuals to be observed at the same times or even the same number of times during the study. Although methods have been developed to allow for discontinuities in the functions of interest and their derivatives (e.g., by repeating knots in a B-spline or with partial linear models (e.g. Shiau et al, 1986; Lee and Shiau, 1994)), most analysis of functional data assumes that the latent functions are smooth. This provides more structure in the models and increases the precision and power of the analysis. A classic example of FDA is the analysis of a set of height measurements taken from different individuals and indexed by age. Although each individual's height can only be measured at discrete times, the target of inference is the set of continuous (infinite dimensional) growth curves relating height to age for the each individual and for the population at large.

It is appropriate to model the effects of teleconnections on climate through FDA because these phenomena are expected to produce long-term effects rather than rapid changes. Several researchers have already applied FDA methods to climate data. The two introductory texts on FDA by Ramsay and Silverman (Ramsay and Silverman, 2005, 2002) repeatedly used data on climate at several Canadian cities to illustrate the methods they introduce. Meiring (2007) applied FDA to study variations in atmospheric ozone pressure as a function of elevation, relating these changes to several predictors including an index of the quasi-biennial oscillation, solar activity, and time of year. In another study, Temiyasathit et al (2009) modeled functional observations of daily ozone concentration at 14 locations in the Dallas-Fort Worth area over a 3 year period and applied spatial methods to reconstruct concentrations over the entire region. Escabias et al (2005) developed a method of logistic regression with functional predictors and applied it to model the probability of drought in a given year as a function of the temperature profile. In their method, functional principal components is used to reduce the yearly temperature profiles to a small number of predictors. Aguilera et al (2008) extended this method by incorporating an ARIMA time series model of the principal components that accounts for correlations in observations from the same location and examined the relationship between monthly temperature and drought in Granada, Spain. Ocaña Lara et al (2009) described a model for predicting pollen counts given temperature in which both the response and predictor were treated as continuous functions. With regards to modeling teleconnections, Valderrama et al (2002) analysed seasonal time-series of the ENSO index to illustrate a method for computing functional principal components from time series with correlated observations.

The goal of our analysis was to study geographical variations in the climate impacts of the four teleconnections across BC through functional principal components analysis (FPCA). In particular, we focused on modeling the impacts of the teleconnections on mean monthly temperatures and our analysis addressed two key challenges. The first was that fine scale weather data for BC was not available from a single source. Instead, we work with a large data set accumulated from several agencies (as described in Section 2). Although the total number of weather stations contributing data was quite large, the agencies tended to change the configuration of the stations frequently. With the exception of a small number of long-term monitoring stations, the stations were active for short periods of time and the stations' locations changed continually. This made it difficult to model the raw data directly. To address this, we constructed complete time series of pseudo-data at fixed grid points by simulating observations from regression spline models fit to the monthly observations. Multiple pseudo-data sets were generated and analysed together to account for uncertainty in the simulation process. This stage of our analysis is described in Section 3.1. The second challenge was that the impacts of the teleconnections cannot be observed directly. This contrasts with the studies of observed temperatures, ozone pressures, or precipitation levels as described above. Instead, it is necessary to estimate the effects of the teleconnections. We fit regression models to the series of pseudo-observations at each grid point and then conducted FPCA using the estimated teleconnection effects. This stage of the analysis is described in Sections 3.2, 3.3, and 3.4. We summarize the results of our analysis in Section 4 and conclude with further discussion in Section 5.

2 Data

2.1 Climate Data

Reference daily climate data for stations in BC operating between 1872 and 2005 were compiled from Environment Canada's long-term station monitoring network (Canadian Meteorological Service). These data were supplemented with additional records obtained from inter-governmental agencies that operate stations at high elevations in Southern BC watersheds. The data were quality controlled/quality-assured prior to delivery to ensure consistency across the stations and to remove instrumentation artifacts. The data set contains observations for a total of 1602 stations recording daily minimum, maximum, and mean temperatures and/or daily precipitation for at least one day in the 134 year period. The current analysis focuses on modeling monthly mean temperature. Challenges of extending our work to modeling precipitation are discussed in Section 5.

Although the total number of stations was quite large, the stations were spread over a large area and many were active for short periods of time so that data were sparse in both space and time. Plotting the years in which each station was recording temperature showed that no more than 500 stations were active at any one time. Fewer than 50% of the stations produced recordings for more than half the days in each of 10 years or more (see Figure S1 in the Supplementary Materials). Further, the

stations tended to cluster in the most densely populated areas of the province. Spatial coverage was particularly poor in the parts of the province with low population density including the north (above approximately 55° latitude) and in isolated areas like the Chilcotin Plateau (near lat 52.0N, long 124.0W). A good representation of the spatial coverage is provided by the locations of the stations active in 2000 (shown in Figure S2 in the Supplementary Materials).

As noted below, data for the four teleconnection indices were available on a monthly scale from 1950 onward. To match this, we restricted our analysis to the 50-year study period from 1951 to 2000 and averaged records of daily mean temperature by month. A total of 1125 stations reported daily temperatures at some point in this period, though the vast majority were active for only a few years. The final climate data used in our analysis consisted of intermittent average monthly mean temperatures over a 600 month period from each of these 1125 stations.

2.2 Teleconnection Indices

Historical records of the four teleconnection indices considered in the analysis were obtained from two separate sources. Monthly values of the standardized SOI, the NAO index, and the PNA index were retrieved from the web-site of the National Oceanic and Atmospheric Administration's (NOAA) National Weather Service Climate Prediction Center [1]. Monthly values of the Pacific Decadal Oscillation Index (PDO) were retrieved from the web-site of Dr. Nathan Mantua in the Department of Atmospheric Sciences at the University of Washington [2]. Further descriptions of these indices are available from these sources.

Initial analysis of the data indicated that the PNA and NAO vary much more rapidly than the SOI and PDO which have longer sustained periods above or below their normal values (traceplots are provided in Figure S3 of the Supplementary Materials). Correlations amongst the 4 indices ranged from -.37 between the SOI and PDO, with 95% confidence interval (-.44,-.30), to .38(.30,.44) between the PDO and PNA. This clearly indicated that the indices were not independent but the correlations were not sufficient to raise concerns about mulitcollinearity.

2.3 Notation

The following notation is used to refer to different components of the observed data:

[1] (`http://www.cpc.noaa.gov/`)

[2] (`http://www.atmos.washington.edu/~mantua/`)

T_{it}	mean temperature (degrees Celsius) observed at station i during month t of the study period ($i = 1,\ldots,1125$, $t = 1,\ldots,600$)
(x_i, y_i)	location of station i measured in Universal Transverse Mercator (UTM) coordinates
z_i	elevation of station i in meters
$w_l(t)$	value of the l^{th} teleconnection index during interval t of the study period ($l = 1,\ldots,4$)

Notation pertaining to the fitted models is introduced as the methods are described in Section 3.

3 Methods

The objective of this analysis was to examine variations in the regional effects of the four teleconnections on mean monthly temperatures across BC over the study period. To do this, it was necessary to estimate the effects of the teleconnections, since these effects cannot be observed directly. One major challenge in our analysis was the large number of stations combined with the relatively small number of observations from each station. This made it difficult to apply our methods directly to the observed data. Instead, we have developed a preprocessing technique to impute complete time series at a select set of fixed points throughout the regions in which stations were active in most years (Section 3.1). Regression models including lagged values of the teleconnection indices to incorporate our assumption of smoothness were then fit to the complete data series as described in Section 3.2. Finally, we applied FPCA to point estimates of the realized effects as described in Section 3.4). Many imputed data sets were constructed and analysed together to correctly account for uncertainty in both the preprocessing and model fitting steps.

3.1 Preprocessing

Data preprocessing was conducted to map the raw observations onto a set of fixed point locations in order to construct complete data series. This removed complications caused by the intermittent activity and changing spatial distribution of the original stations. Fixed points were placed at the vertices of a 50 km square lattice with the origin at Vancouver International Airport (lat 49.2N, long 123.2W). To avoid extrapolation, only locations on the lattice which had at least one active station within a 50 km radius in all 50 years were retained – a constraint that was satisfied by 168 of the lattice points (Figure S2 in the Supplementary Materials identifies these points and illustrates their relation with the stations active in 2000). We refer to the 168 lattice points retained for the analysis as grid points and to the elements of the simulated data at these points as pseudo-data.

Pseudo-data was constructed separately for each of the 600 months in the study period. Models were fit to the observed data in each month and new mean temperature

values were simulated at the 168 grid points accounting for both uncertainty in the fitted models and residual error. The effects of location and elevation on mean monthly temperature were modeled using thin plate regression splines (TPRS) (Wood, 2003). As described in Wood (2003), TPRS are low-rank smoothers obtained by truncating the basis of standard thin-plate splines which require less computation to fit to large data sets. Similar methods are employed in the ANUSPLIN model which is commonly used to interpolate climate data over large areas (Hutchinson, 1995). Further examples of the use of thin plate (regression) splines in modeling or interpolating climate data include Hancock and Hutchinson (2006), Haylock et al (2008), and Xu et al (2009).

Let $\mathscr{A}_t$ denote the set of stations active in the t^{th} month of the study period. The models we fit had the form

$$T_{it} = g_t^{(S)}(x_i, y_i|\gamma_t^{(S)}) + g_t^{(E)}(z_i|\gamma_t^{(E)}) + \varepsilon_{it}, \quad i \in \mathscr{A}_t \tag{1}$$

where $g_t^{(E)}(z_i|\gamma_t^{(S)})$ and $g_t^{(S)}(x_i, y_i|\gamma_t^{(S)})$ are, respectively, 1 and 2 dimensional TPRS with parameters $\gamma_t^{(S)}$ and $\gamma_t^{(E)}$ which model the effects of elevation and location. The variance of the errors in a single time month, $\text{Var}(\varepsilon_{it}) = \sigma_t^2$, was assumed to be the same at all locations and elevations. These models were fit in `R` using the package `mgcv`, and the default knot selection algorithm produced TPRS with 9 parameters for modeling the effects of elevation and 29 parameters for modeling the effects of location. All parameters were allowed to vary independently with time period, t, and inference was conducted by considering penalized least-squares as an empirical Bayes procedure with the smoothing parameters selected by the generalized cross-validation (GCV) criterion (for further details see Wood, 2006, pg. 189-196).

After fitting these models, we generated complete series of pseudo-observations at the 168 grid points by simulating data from the monthly posterior predictive distributions. For each t we first generated values of the parameters $\gamma_t^{(S)^*}$, $\gamma_t^{(E)^*}$, and $\sigma_t^{2^*}$ from the joint posterior distribution of the TPRS model for each $t = 1, \ldots, 600$. We then simulated independent pseudo-observations, T_{it}^*, at each of the grid point conditional on these values:

$$T_{it}^*|\gamma_t^{(S)^*}, \gamma_t^{(E)^*}, \sigma_t^{2^*} \sim N\left(g_t^{(S)}(x_i^*, y_i^*|\gamma_t^{(S)^*}) + g_t^{(E)}(z_i^*|\gamma_t^{(E)^*}), \sigma_t^{2^*}\right), \quad i = 1, \ldots, 168.$$

Here (x_i^*, y_i^*) represents the location of the i^{th} grid point in UTM coordinates and z_i^* represents the estimated of the elevation obtained from digital elevation models, $i = 1, \ldots, 168$. As we discuss in Section 3.3, randomly generating pseudo-observations rather than computing fitted values based on point estimates of the TPRS parameters allowed us to account for both uncertainty in the fitted surfaces and residual error when fitting our regression models.

3.2 Regression Analysis

In the second stage of our analysis, we fit separate linear regression models to the pseudo-observations generated at each of the 168 grid points. The models combined the effects of the four teleconnection indices with linear trends and seasonal effects.

Initial analysis of the data indicated that the variance of residuals differed significantly over space and time, both within and between years. In particular, the pseudo-observations seemed to have higher variability in the early years when fewer stations were active, in the winter when temperature is less stable, and in the north where data was more sparse. Further modeling of the error variance was included to account for these effects.

The explicit model for the pseudo-observation at grid point i in month t at had the form

$$T^*_{it} = \beta_{0i} + \beta_{1i} \cdot \text{year}(t) + s_i(\text{month}(t)) + \sum_{l=1}^{4} f_{li}(t) + \varepsilon^*_{it} \tag{2}$$

where the functions $\text{year}(t) = \lceil t/12 \rceil$ and $\text{month}(t) = \mod(t,12)$ denote the year of the study and corresponding month within the year. The term β_{0i} represents the intercept (the overall mean temperature in the first year of the study), β_{1i} a linear term across years, and $s_i(\text{month}(t))$ a seasonal effect that is constant across years. The last effect was modeled with a cyclic cubic regression spline. As described by Wood (2006, pg. 151–152), a cyclic cubic regression spline is a cubic regression spline defined over some interval with the added restrictions that the value of the function and its first two derivatives be equal at the interval's end points. In our model, $s_i(\cdot)$ is defined over the range $(0,12)$ so that $s_i(0) = s_i(12)$ and the first two derivatives are also equal at these points. This ensures that the seasonal effect is smooth over the entire year – even between December ($s_i(12)$) and January ($s_i(1)$). Cyclic cubic regression splines can be fit by using a standard cubic regression spline basis and placing restrictions on the spline coefficients or by computing modified basis functions. In fitting our models, we used the basis of modified cubic regression spline basis functions as computed with the `mgcv` package using the default placement of knots. This produced a basis of 8 functions, and we denote the vector of parameters for this function by $\boldsymbol{\beta}_{2i}$.

The final term in this equation (2) models the effects of the four teleconnection indices. Specifically, $f_{li}(t)$ models the effect of the l^{th} teleconnection on mean temperture at grid point i in month t. As noted in the introduction, we believed *a priori* that the effects of these predictors should be smooth over time. To account for this belief, we allowed for a lag of up to 24 months so that the effect is defined as

$$f_{li}(t) = \sum_{d=0}^{24} \beta_{3li}(d) w_l(t-d).$$

and fit the model only to the last 576 months of pseudo-data ($t = 25,\ldots,600$). Further to this, we modeled the vectors of coefficients for each teleconnection, $\boldsymbol{\beta}_{3li} = (\beta_{3li}(0),\ldots,\beta_{3li}(24))'$, $l = 1,\ldots,4$, as points along a smooth curve. Specifically, we modeled the elements of $bm\boldsymbol{\beta}_{3li}$ points along a cubic regression spline implemented using the Bayesian P-spline approach of Lang and Brezger (2004). This framework provides a Bayesian implementation of the penalized spline (P-spline) methods of Eilers and Marx (1996). Each curve was modeled as the linear combination of B-spline basis functions and smoothness was imposed by assigning a hierarchical prior that favored small differences between adjacent spline coefficients. Specifically, we modeled $\boldsymbol{\beta}_{3li} = \boldsymbol{B}\boldsymbol{b}_{3li}$ where $\boldsymbol{B}$ is a matrix of cubic B-splines evaluated at the points

$0,\ldots,24$. The vector of coefficients $\boldsymbol{b}_{3li}$ was then assigned a prior based on first order differences so that

$$b_{3li}(j) - b_{3li}(j-1) \sim N(0, 1/\lambda_{li}^2), j = 1,\ldots,24$$

and $b_{3li}(0)$ was assigned a vague normal prior with large variance.

The parameter λ_{li}^2 in this model is equivalent to the smoothing parameter in classical spline methods and controls both the complexity of $\boldsymbol{b}_{3li}$ and the smoothness $f_{il}(t)$. If λ_{li}^2 is big then the elements in $\boldsymbol{b}_{3li}$ will all be similar and $f_{li}(t)$ will be smooth. In the extreme case, the elements of $\boldsymbol{b}_{3li}$ will be almost exactly equal so that $b_{3li}(d) \approx b_{3li}(0)$ for all $d = 1,\ldots,24$. Then $\boldsymbol{\beta}_{3li} = \boldsymbol{B}\boldsymbol{b}_{3li} \approx b_{3li}(0)\boldsymbol{B}\boldsymbol{1} = b_{3li}(0)$, and the overall effect of the teleconnection in month t will be proportional to the average of the values of teleconnection index over the previous 25 months ($f_{li}(t) = \sum_{d=0}^{24} \beta_{3li}(d) w_l(t-d) \approx b_{3li}(0) \sum_{d=0}^{24} w_l(t-d)$). If λ_{li}^2 is small then the elements of $\boldsymbol{b}_{3li}$ will vary widely and $f_{li}(t)$ may be very different from one month to the next. Following Lang and Brezger (2004), we have assigned λ_{li}^2 a gamma prior such that $1/\lambda_{ij}^2$ has a small expected value and large variance (see Appendix A).

In our analysis, we selected a basis with 25 functions built from equally spaced knots. The resulting design matrix, $\boldsymbol{B}$, has dimension 25×25, so that the spline is of full rank. This implies *a priori* that $\boldsymbol{b}_{3li}$ can take any value in $\Re^{25}$, though the prior distribution favors those vectors which are smooth.

The residuals errors, ε_{it}^*, were modeled as independent, mean 0 normal random variables with location and time dependent variance given by

$$\mathrm{Var}(\varepsilon_{it}^*) = \tau_i^2 \exp\left(\rho_{1i} \cdot \mathrm{year}(t) + v_i(\mathrm{month}(t))\right).$$

In this model, τ_i^2 denotes the base variance at grid point i, ρ_{1i} represents a yearly linear term accounting for the decrease in variance over time, and $v_i(t)$ represents a second cyclic spline modeling the seasonal changes in the residual variance. As for $s_i(t)$, we modeled $v_i(t)$ using the 8 modified cubic regression spline basis functions produced by the default knot placement in the `mgcv()` package, and we denote the vector of coefficients for $v_i(t)$ by ρ_{2i}. The full model at each grid point contains 10 free parameters and 100 penalized parameters in the mean function and a further 10 free parameters in the variance function to be estimated from $600 - 24 = 576$ pseudo-data points.

3.3 Inference and Multiple Imputation

Inference for these models was computed separately at each of the 168 grid points. To account for uncertainty in the simulation of the pseudo-observations, we generated 100 independent sets of pseudo-data as described in Section 3.1. Bayesian methods were then applied to fit the model described in Section 3.2 and to combine the results from the different pseudo-data sets. Markov chain Monte Carlo (MCMC) was used to simulate from the posterior distribution of the model parameters at each of the 168 grid points conditional on each of the 100 pseudo-data sets. We found that the individual Markov chains converged quickly, and each chain was run for 3000

iterations. The first 1000 iterations were discarded as burn-in and the final 2000 iterations, thinned every 20 iterations to save storage space, were retained. This produces a combined sample of 10,000 iterations for computing posterior summary statistics at each of the 168 grid points. Prior distributions were selected to be non-informative, except for $\boldsymbol{b}_{3li}$ as specified above. Details are provided Appendix A.

Our approach of combining inference from multiple simulated data sets is akin to the multiple-imputation methods of Little and Rubin (2002). These methods are often applied in the analysis of data with missing values. Imputation models are used to construct complete data sets and variance estimates are computed that combine uncertainty in both the simulated data and in the model parameters given the complete data. Our analysis essentially treats the pseudo-data as the complete data and ignores the observed data because of the difficulties described above. Approximate formulas like those given by Little and Rubin (2002, pg. 211) could be used to compute variance estimates and to construct approximate credible intervals. The advantage of combining samples from the 100 separate posterior distributions generated by MCMC is that interval estimates can be computed without distributional assumptions (e.g., approximate normality).

3.4 Functional Principal Components Analysis

In the final stage of our analysis, we applied FPCA to examine variations in point estimates of the effects of the teleconnections across the grid points. Standard multivariate PCA provides a way to extract the principal modes of variation from a set of vectors and to approximate the vectors as linear combinations of a small number of these factors. Functional principal components analysis extends these methods to examine the variation in a set of functions (vectors of infinite length). We begin with a brief description of multivariate PCA, and follow with the extension to FPCA and the specific methods applied in our analysis.

The objective of multivariate PCA is to reduce the dimensionality in a set of observed vectors. Let $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_N$ be a set of vectors of length p. The first principal component is defined by finding the vector, $\boldsymbol{\psi}_1$, which maximizes the variance of the inner products

$$\xi_{i1} = \langle \boldsymbol{v}_i, \boldsymbol{\psi}_1 \rangle = \sum_{j=1}^{p} v_{ij}\psi_{1j}, i = 1,\ldots,N$$

subject to the constraint that $\boldsymbol{\psi}_1$ has unit length, $\langle \boldsymbol{\psi}_1, \boldsymbol{\psi}_1 \rangle = 1$. Subsequent components are defined sequentially by finding the vector, $\boldsymbol{\psi}_k$, which maximizes the variance of the values

$$\xi_{ik} = \langle \boldsymbol{v}_i, \boldsymbol{\psi}_k \rangle = \sum_{j=1}^{p} v_{ij}\psi_{kj}, i = 1,\ldots,N$$

subject to the constraints that $\boldsymbol{\psi}_k$ have unit length and be orthogonal to the previously defined vectors:

1. $\langle \boldsymbol{\psi}_k, \boldsymbol{\psi}_k \rangle = 1$, and

2. $\langle \psi_k, \psi_m \rangle = 0$ for every $m < k$.

The vector ψ_k is the k^{th} loading vector and ξ_{ik} is the k^{th} principal component for sample vector i (also called the k^{th} principal component score). The loading vectors are equivalent to the eigenvectors of the sample covariance matrix of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$, and the number of unique components that can be extracted is limited by the rank of this matrix and is always less than or equal p (see e.g. Jolliffe, 2002, Chapter 3).

In practice, it is common that a small number of components, $K << p$, explain most of the variance in the original vectors. In this case, $\boldsymbol{v}_i$ may be approximated by the truncated sum

$$\boldsymbol{v}_i \approx \sum_{k=1}^{K} \xi_{ik} \psi_k.$$

This approximation provides an alternate interpretation of the loading vectors and the principal components. For any $k \leq p$, $\psi_1, \ldots, \psi_k$ is the set of orthonormal vectors that produces the best linear approximation to the original data. The principal components are the coefficients in these linear approximations.

Functional PCA applies a similar methodology to reduce the dimension of a sample of N functions, $f_1(\cdot), \ldots, f_N(\cdot)$, defined over some range $\mathscr{S} \subset \Re$. The essential difference is that the inner product between two functions, say $f(\cdot)$ and $\psi(\cdot)$, is defined by the integral

$$\langle f, \psi \rangle = \int_{\mathscr{S}} f(t)\psi(t)\, dt$$

in place of the sum above. Given this definition of the inner product, the first principal component is defined by finding the function, $\psi_1(\cdot)$, which maximizes the variance of the inner products $\xi_{i1} = \langle f_i, \psi_1 \rangle$, $i = 1, \ldots, N$, subject to the constraint $\langle \psi_1, \psi_1 \rangle = 1$. Subsequent components are defined sequentially by finding the functions $\psi_k(\cdot)$ which maximize the variance of the values $\xi_{ik} = \langle f_i, \psi_k \rangle$, $i = 1, \ldots, N$, subject to the constraints

1. $\langle \psi_k, \psi_k \rangle = 1$, and
2. $\langle \psi_k, \psi_m \rangle = 0$ for every $m < k$.

As in standard PCA, it is common that the first few components explain the majority of the variance and that each sample function may be approximated by a truncated sum

$$f_i(t) = \sum_{k=1}^{K} \xi_{ik} \psi_k(t), \quad t \in \mathscr{S}$$

for some small K. Ramsay and Silverman (2005) refer to the functions $\psi_1(\cdot), \ldots, \psi_K(\cdot)$ as the principal component curves or harmonics and to the values ξ_{ik}, $i = 1, \ldots, N$, $k = 1, 2, \ldots, K$, as the principal components or principal component scores.

Several methods are available to conduct FPCA for a set of functions, and we have applied the methods of Ramsay and Silverman (2005, Section 8.4.2). In practice, the functions can only be observed at discrete time points, and so it is necessary to recover the underlying functions from the vectors of observations. Let $\boldsymbol{v}_i$ be the vector of possibly noisy observations of $f_i(\cdot)$ at times $t_{i1}, \ldots, t_{p_i}$ (note that the time points and number of observations may vary among the samples). In the method of Ramsay

and Silverman (2005), $f_i(\cdot)$ is approximated by modeling $\boldsymbol{v}_i$ as a linear combination of basis functions, $\phi_1(\cdot),\ldots,\phi_R(\cdot)$, so that

$$v_{ij} \approx \sum_{r=1}^{R} c_{ir}\phi_r(t_j) = \boldsymbol{\phi}(t_j)'\boldsymbol{c}_i, \quad t = t_{i1},\ldots,t_{p_i i}$$

where $\boldsymbol{\phi}(t_j) = (\phi_1(t_j),\ldots,\phi_R(t_j))'$ and $c_i = (c_{i1},\ldots,c_{iR})'$. The i^{th} sample function is then taken to be equal to $f_i(t) = \boldsymbol{\phi}(t)'\boldsymbol{c}_i$ for any $t \in [t_1,t_p]$. Following this, Ramsay and Silverman (2005, pg. 162–163) show that FPCA for the functions $f_1(\cdot),\ldots,f_N(\cdot)$ can be computed from the matrix of spline coefficients, $C' = (\boldsymbol{c}_1,\ldots,\boldsymbol{c}_N)$. Specifically, the k^{th} principal component curve is $\psi_k(t) = \boldsymbol{\phi}(t)'\boldsymbol{b}_k$, $t \in [t_1,t_p]$, where $\boldsymbol{b}_k = W^{-1/2}\boldsymbol{u}_k$, W is the $R \times R$ matrix formed by the inner products of the basis functions, $W_{ij} = \langle \phi_i, \phi_j \rangle$, and $\boldsymbol{u}_k$ is the k^{th} eigenvector of the matrix

$$M = N^{-1}W^{1/2}C'CW^{1/2}.$$

The k^{th} principal component for the i^{th} observation is $\xi_{ik} = \langle f_i, \psi_k \rangle = \boldsymbol{c}_i'\boldsymbol{W}\boldsymbol{b}_k$.

In our analysis of the BC climate data, we applied this method of conducting FPCA to study variations in point estimates of the effects of the four teleconnections among the grid points over the study period. Letting $\hat{\boldsymbol{\beta}}_{3li}$ denote the posterior mean of $\boldsymbol{\beta}_{3li}$, we computed point estimates of the effect of the l^{th} teleconnection at grid point i by

$$\hat{f}_{li}(t) = \sum_{d=0}^{24} \hat{\boldsymbol{\beta}}_{3li}(d) w_l(t-d), \quad t = 25,\ldots,600.$$

We then recovered the functions underlying the vectors $\hat{\boldsymbol{f}}_{l1},\ldots,\hat{\boldsymbol{f}}_{l,168}$ by approximating each vector as a linear combination of cubic B-splines with 150 equally spaced knots. Note that Ruppert et al (2003) recommend using the minimum of 35 and $p/4$ knots. However, we found that 35 knots was not sufficient to capture the variations in the teleconnection effects. Instead, we opted for the higher bound and regularized the curves by penalizing the integrated second derivative choosing the smoothing parameter by minimizing the GCV. Functional PCA of the resulting functions $f_{l1}(\cdot),\ldots,f_{l,168}(\cdot)$ was then performed by applying the method of Ramsay and Silverman (2005) as described above. This analysis was conducted separately for each of the teleconnection indices (i.e., for each $l = 1,\ldots,4$) and implemented using the `fda` package in `R` as described in Ramsay et al (2009).

4 Results

4.1 Preprocessing

The models fit in the preprocessing step were validated by simulating new data at select stations and comparing this with the observed data. Plots comparing the simulated and observed values in 1951, 1976, and 2000 at five stations spanning the range of latitudes are provided in the supplementary materials (Figure S3). The mean

squared error for the simulated values was less than 1.7 °C [2] and the range of the simulated values covered the observed values in all cases. We found no evidence that the accuracy of the TPRS models for simulating the pseudo-data was affected by either spatial location, elevation, or year.

Examples comparing the pseudo-data generated at the fixed grid points and the data observed at nearby stations are also provided in the supplementary materials (Figure S4). In these plots, we have compared the data simulated at five grid points in the years 1951, 1976, and 2000 with observed data at two nearby stations – the closest station active in that year and the closest active station within 100 m elevation. As expected, there is good agreement between the observed data and the simulated pseudo-data when the grid point was close to an active station at similar elevation (e.g., grid point 26 in all three years). Not surprisingly, differences between the observed data and the pseudo-data are large if there was a large distance or difference in elevation between the grid point and the nearest station (e.g, grid point 101).

4.2 Regression Modeling

In this section we summarize our inference for the simple trends, seasonal trends, and variance components for the 168 grid points. Point estimates, standard deviations (SD), and interval estimates provided refer to the posterior means, posterior standard deviations, and 95% credible intervals (CIs) computed from the combined MCMC samples as described in section 3.3. Note that our analysis cannot provide variance or interval estimates for estimates for quantities that combine information from multiple grid points (e.g., average effects) because it does not yet incorporate spatial correlation. Extensions to incorporate spatial correlation are discussed further in Section 5. Overall, we found that the regression model in equation (1) explained between 86.0 and 94.4% of the variation in the pseudo-data at each of the 168 grid points.

4.2.1 Simple Trends

Point estimates of the grid point specific intercept parameters, β_{0i}, showed a clear spatial pattern (see Figure 1 – Top Left Panel). As expected, estimates of β_{0i} were generally highest in the southeast of the province and along the Pacific coast, and lowest along the eastern border and in the north of the province. The largest point estimate for β_{0i} was 10.08 °C (SD .09 °C) at lat 49.2 N, long -121.8 W in the Fraser Valley near Chilliwack, BC. The smallest point estimate was -1.52 °C (SD .14 °C) at lat 58.0 W, long -130.0 N, a grid point in the far north of the province. Posterior standard deviations for β_{0i} ranged from .08 °C to .16 °C.

Estimates of the linear term, β_{1i}, were all greater than zero indicating a consistent warming trend across the entire province (see Figure 1 – Top Right Panel). The mean of the 168 point estimates of β_{1i} was .22 °C/decade, corresponding to an average increase of 1.02 °C over the 48 year study period. Posterior standard deviations for β_{1i} were all between .07 and .12 °C/decade, and the lower bound of the 95% CI for β_{1i} was greater than 0 °C at 48 of 168 (28.6%) of the grid points.

Posterior standard deviations for both $\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$ tended to be larger in the east and north of the province (see Figure 1 – Bottom Panels). This pattern directly reflected the spatial pattern in the estimates of the variance parameter, τ_i^2, as discussed below.

4.2.2 Seasonal Trends

As expected, seasonal trends at all grid points showed a clear and significant increase in mean monthly temperature during the summer and decrease in the winter. A plot of the seasonal effect averaged over all grid points is provided in Figure 2. Overlayed on this plot are the estimated seasonal trends for the grid points with the largest and smallest seasonal differences. The difference between the January and July temperatures averaged over all of the grid points was 20.51 °C. The largest seasonal effect was found at lat 58.6 W, long -121.5 N in the northern Rocky Mountains and spanned a difference of nearly 36.66 °C (SD=.71) between January and July. The smallest seasonal effect was found at lat 49.6 W, long -126.6 N on the west coast of Vancouver Island and spanned a difference of only 9.96 °C (SD=.34) between January and August. As with the simple trends, the estimates were more precise in the southwest of the province where temperatures were more stable and more information was available.

4.2.3 Variance Structure

Point estimates of the base variance parameter, τ_i^2, at the 168 grid points varied widely from 1.78 $°C^2$ at lat 49.6 W, long -126.6 N to 7.63 $°C^2$ at lat 58.63 W, long -121.50 N. Point estimates were generally higher in the north and east of the province and lower in the south and along the west coast (see Figure S5). This pattern likely related to both systematic changes in the variance and differences in the experimental error. Temperatures in the south and along the coast were more stable than in the north and east of the province. Stations in the north and east were also more widely spaced so that there is less information at these grid points. Point estimates of the linear term in the variance model, the ρ_{1i}'s, were below 0 at 153 of the 168 grid points (91.1%). However, the point estimates were all less than .01 °C/year in magnitude and the corresponding 95% CIs covered 0 at all 168 stations. It is not possible to conclude that any that a consistent change in variance occurred at any of the stations. A plot of the estimated seasonal affect on the variance ($\exp(v_i(t))$) averaged over all grid points is provided in Figure S6. Overlayed on this plot are the estimated seasonal trends for the grid points with the largest and smallest seasonal changes in variance. The largest seasonal change was found at lat 55.9 W, long -120.8 N where the variance of the residuals was estimated to be 17.78 times higher in the winter than in the summer. The smallest seasonal change was found at lat 50.0 W, long -126.7 N where the estimated difference was only 2.7 times. On average, the variance of the residuals was 6.37 times higher in the winter than in the summer.

4.3 Multiple Imputation

To assess the effect of multiple imputation on the inference from our regression models, we estimated the proportion of variance arising between replicates of the pseudo-data. That is, the proportion of variance in the posterior distribution of each parameter that can be attributed to variation in the pseudo-data as opposed to uncertainty in the parameters given a specific pseudo-data set. Let θ_i denote any parameter in the regression model of equation (2). We estimated the percent of the posterior variance for θ_i arising from variations between replicates of the pseudo-data by

$$100\left(\frac{\mathrm{Var}(\hat{\theta}_i)}{\mathrm{Var}(\hat{\theta}_i)+\mathrm{E}(\widehat{\mathrm{Var}}(\theta_i))}\right)$$

where $\mathrm{Var}(\hat{\theta}_i)$ and $\mathrm{E}(\widehat{\mathrm{Var}}(\theta_i))$ represent the variance in the posterior means and the mean of the posterior variances over the 100 pseudo-data sets. Figure S14 presents plots of these values for the intercept (β_{0i}), the linear term (β_{1i}), and the leading coefficients of the teleconnection effects ($\beta_{3li}(0)$, $l = 1,\ldots,4$). Estimates of the percent of variance between replicates ranged, on average, from 8.52% for the leading coefficient of the SOI to 27.61% for the leading coefficient of the PNA. This indicates that we would have underestimated the uncertainty in these parameters by as much as 28% if we had generated a single pseudo-data set.

4.4 Impacts of Teleconnections

Finally, we describe the results of the FPCA of the estimated effects of the four teleconnections as described in Section 3.4. To assess the importance of the four teleconnection indices in the final model, we computed the mean square of the estimated effects at each station (i.e., $\sum_{t=25}^{600}\hat{f}_{il}(t)/576$ for each $i = 1,\ldots,168$ and $l = 1,\ldots,4$). Mean squares for the estimated effect of the PNA were consistently largest, .34 $^\circ\mathrm{C}^2$ on average with a range of .09 $^\circ\mathrm{C}^2$ to .69 $^\circ\mathrm{C}^2$. The average mean square of the estimated effects for the PDO was .03 $^\circ\mathrm{C}^2$ (range .01 to .12 $^\circ\mathrm{C}^2$), and the average mean square of the estimated effects for the SOI was .04 $^\circ\mathrm{C}^2$ (range .01 to .10 $^\circ\mathrm{C}^2$). This suggests that these predictors produced effects of similar magnitude. Mean squares for the NAO were consistently small, .01 $^\circ\mathrm{C}^2$ on average (range .00 to .06 $^\circ\mathrm{C}^2$) indicating that this teleconnection had the smallest effect on monthly mean temperature. We report our results in this order.

4.4.1 PNA

Functional PCA of the realized effects for the PNA indicated that the first two components explained 81.2% and 17.0% of the variability between grid points, for a total of 98.2%. Plots of the mean function and the first two principal component curves are provided in Figure 3. In these plots, the red and blue segments highlight periods when the curves were above or below 0 $^\circ$C for at least 12 consecutive months. Plots of the scores for the first two principal components are provided in Figure 4.

The mean function suggests that the PNA affected the mean monthly temperature averaged over all grid points by as much as -.40 °C, both positive and negative. As expected, the mean effect of the PNA at the 168 grid points was semi-periodic and moved back and forth across 0 °C several times over the 50 year period. However, there were several periods of 12 months or more during which the mean function was consistently above or below 0 °C. These periods appear to correspond to known periods of extreme temperature (see Section 5).

Scores for the first functional principal component show a clear spatial effect with large positive values in the northwest of the province and large negative values in the southeast. Figure 5 (Top Panel) illustrates the effect of this mode of variation by comparing the mean function (in black) with the mean plus and minus the first principal component curve times one standard deviation of the scores (in red and blue respectively). Most importantly, the first component appears to describe a phase shift in the effect of the PNA. This is shown more clearly in Figure 6 which provides the same information for the last 10 years of the study period. The impact of PNA appears to arrive earlier where the scores are negative (in the southeast of the province) and later where the scores are positive (in the northwest of the province). Further, the plot shows that the peaks and valleys were often, though not always, more pronounced in the southeast of the province where the scores were negative.

Scores for the second functional principal component also show a clear spatial pattern with the largest values in the center of the province. Plots of the effect of this component indicate that it affected the magnitude of the realized effect of the PNA (see the bottom panels of Figures 5 and 6). The estimated effect of the PNA was stronger in the center of the province where the scores are positive and weaker in the north and along the west coast where the scores are negative.

4.4.2 PDO

Functional PCA of the realized effects of the PDO indicated that the first two components explained 56.0% and 39.4% for the variability between grid points, for a total of 95.4%. Plots of the mean effect, the first two principal component curves, and the corresponding scores are provided in the supplementary materials (Figures S7 and S8).

The mean function suggests that the PDO affected the mean monthly temperature averaged over all grid points by as much as -.36 °C or .31 °C. The mean of the realized effects of the PDO still oscillated over the study period, but seemed to remain negative in most months up to 1975 and positive in most months after 1980. This suggests that the PDO had a negative impact on monthly temperatures at most stations in the first half of the study period and a positive impact in the second half of the study period.

Scores for the first two functional principal components again showed clear spatial patterns. Figure S9 illustrates the effects of these two modes of variation on the impact of the PDO. The first principal component curve takes values above 0 °C for most of the months up to 1975 and values below zero thereafter. This accentuated the difference in the impact of the PDO between the first and second half of the study periods in the north of the province where the scores were negative or lessened the

difference in the south of the province where the scores were positive. The second principal component curve appears to increase or decrease the impact of the PDO – maintaining the overall shape of the mean but increasing or decreasing the magnitude of the oscillations. The impact of the PDO was smoother in the east where the scores are positive and oscillated more in the west where the scores are negative.

4.4.3 SOI

Analysis of the estimated effects of the SOI indicated that the first two components explained 84.5% and 13.6% of the variability between grid points. Plots of the mean effect, the first two principal component curves, and the corresponding scores are provided in the supplementary materials (Figures S10 and S11).

Heuristically, the impact of the SOI is similar to that of the PNA producing several long periods when the average effect was positive or negative. Figure S12 illustrates the effects of the first two principal components on the mean function. Again, the effects of the principal component curves are qualitatively similar to those from the PNA. The first principal component curve seems to represent a phase shift in the effect of SOI and the second seems to represent a change in magnitude. Interestingly, the scores for the first component of the SOI and PNA are of opposing sign so that the impact of the SOI arrives earlier in the north west of the province where scores were negative and later in the south west of the province where the scores were positive. Spatial effects of the second component of variation were less clear, though scores were generally negative – representing a lessened effect – in the south central region of the province.

4.4.4 NAO

Analysis of the realized effects of the NAO indicated that the first component explained 96.1% of the variability between grid points. Plots of the mean effect, the first two principal component curves, the corresponding scores, and the effects of the principal component curves are provided in the supplementary materials (Figures S13, S14, and S15).

Point estimates for the mean function were very close to 0 – less than .07 °C in absolute value for all months. This suggests that there was very little effect of the NAO on mean monthly temperature when averaged over all 168 grid points.

5 Discussion

The goal of this research was to study spatial variations in the impacts of multiple, concurrent teleconnections across BC through methods of FDA, specifically FPCA. We focused attention on mean monthly temperature in this analysis as an indicator of the overall effect of temperature on crop suitability. Further measures like minimum and maximum monthly/daily temperature could be examined as limiting factors in the range of crop suitability due to cold (stress) tolerance (threshold and duration) or growth potential (e.g., growing degree days).

Modeling spatial variations in the impacts of teleconnections presents several challenges. In this research we have developed methods to work with intermittent weather records accumulated from different sources and to estimate the impacts of teleconnections on mean monthly temperature prior to conducting FPCA. Remaining challenges include accounting for spatial correlations in the FPCA, incorporating seasonal differences in the impacts of the teleconnections, and extending our models to examine the impacts of the teleconnections on precipitation levels. We first discuss the interpretation and implications of our findings linked to climate and agricultural knowledge and relevant findings for Canada and BC previously reported in the literature and then discuss future extensions of our methods.

The estimated rate of annual mean temperature increase from our analysis supports a average warming trend across the entire province of approximately 1.02 °C over the 48 year study period. This estimate is based on our data set of 1125 stations, 1951-2000 and includes many higher elevation stations, and is consistent with a previously reported annual mean temperature increase across Canada as a whole of 1.4 °C. This value was derived from 210 high-quality stations with 7 years of additional data, i.e. 1950-2007, whereby strongest warming was reported over western and northwestern Canada of values greater than 1.5 °C, and ranging between 1.5-3 °C (Environment Canada, Canadian Climate Trends, Climate Research Division). The deviation between our value and these other estimates likely relates to differences in the elevation of the stations, station temporal and spatial coverage, total number of records included, and possibly differences in quality-control correction. A warming trend in mean temperature across the province implies benefits for the growth of annual crops within a changing duration of growing season; however, the associated changes in cold stress may imply reduced survival for perennial crops (highly-valued fruit crops and other woody plants) which require cold temperatures to break dormancy and to enable transitioning in their phenology (i.e., further crop development and growth). Our grid-based maps based on historical observations are useful for decision-making regarding crop suitability, but further examination in connection with high-resolution climate scenario-derived trends is required. Further analysis is warranted on how fine-scale differences (i.e., at the micro-climate scale) of mean temperature might explain spatial variance in crop suitability observed from other independent plant/crop phenological survey sampling records.

Our findings reveal marked differences in the relative magnitude in the monthly temperature effects associated with spatial impact of PNA, PDO, ENSO/SOI and NAO. In our models, PNA exhibited the strongest teleconnection-linked variation in mean temperature, higher than both the PDO and SOI/ENSO, with NAO anomoly variation relatively flat. Our findings show higher variation of mean temperature with PNA anomoly for northwestern BC, with marked lower variation to the southeast of the province. This finding is consistent with those of Sheridan (2003), that the PNA has an extensive influence along the northern Pacific coast based on variation of synoptic weather-type frequencies to different teleconnection phases.

In contrast to findings reported by Sheridan (2003) that a strong PNA pattern is often observed as a response to a warm event that resembles ENSO variability, we find the anomoly pattern for SOI/ENSO shows low variation northward, and higher variation to the southward and showed the inverse of the PNA variation pattern. In-

stead, our findings show a stronger similarity between PDO and SOI/ENSO, which would be expected as they are both SST-driven indices, with PDO well-known as a long-lived ENSO like pattern of variation. Therefore, our finding regarding close association of PDO and ENSO patterning is consistent with current statistical climatological knowledge. Wang et. al., (2006) have investigated the influence of Pacific climate patterns (PDO, ENSO) on river flow response across BC and the Yukon based on stream discharge data from hydrometric stations across each of four sub-regions of BC (i.e., southern coast (SC), southern interior (SI), northern coast (NC), northern interior(NI)). Coherent responses were identified between PDO and ENSO with PDO influencing low-flows more significantly and consistently than the ENSO signal; strong positive associations across sub-regions SC, SI and NI where high frequencies and low magnitudes of low-flows were associated with warm PDO phases, and negative associations in NC, where high frequencies and low magnitudes of low-flows were associated with cool PDO phases. This close association between PDO and ENSO is also consistent with our temperature anomoly field findings that show lower variation northward (latitudes lower than Yukon) and stronger variation to the south. We therefore infer the importance of further understanding coupling effects between temperature (minimum, mean, maximum) and precipitation variation in association with changes in elevation, surface snow/ice melt, frontal/Coastal storms and convective storms in the interior of BC. Further supporting our findings here, Shabbar et. al., (1996) has previously reported ENSO-driven impacts on surface and lower tropospheric temperature fields over Canada (500-1000 hPa thickness data) across 1946-1994. They found stronger positive surface temperature anomolies spreading eastward from the west coast, and negative anomolies spreading southeastward from the Yukon, following the onset of ENSO episodes. Clearly, high/low areas in magnitude and their spatial extent of anomoly variation in temperature (and precipitation) alters the conditions for crop growth and survival alongside changes in their response (i.e., stress, sensitivity) for both annual crops during the growing season as well as perennial crops through out the year.

Realized variations in the impact of PNA on mean monthly temperatures identified by our models are consistent with reported extreme event conditions (i.e., summer Pacific Coast/BC Mountain) ranked warmest to coolest across the study period. The period 1958-61 identified by our model was a notable warm episode (positive temperature anomoly/departure) within the top 12 recorded events, that is associated with positive PNA anomoly. Similarly, 1997-99 was another warm episode, with particular warm temperatures in the BC interior, also consistent with an episode of positive anomoly linked with the PNA identified by our model. Many years post-2000 show marked temperature anomolies that rank in the top 12 most extreme years on record. If data were available, extending our analysis to include years up to 2012 would help to lend stronger support for associating the magnitude and duration of episodes of low/high temperature anomoly or precipitation/river streamflow with regional climate teleconnections.

Our findings here across BC showcase the importance of understanding the historical and future, potential spatial impact of multiple teleconnections. Such information may considerably improve spatially-explicit forecasting of future crop potential production and decision-making involving changes in the distribution of cropland under

changing climate variability. Further work aims to extend this analysis, applying the FPCA method across the Prairie provinces (Alberta, Saskatchewan, Manitoba) and associate temperature and precipitation anomoly patterns associated with the teleconnections with historical changes in crop yield and station climate data (historical and more recent near-real time (NRT) records). We anticipate this will require a seasonal-based analysis, as well as adapting our current methodology to account for spatial dependence/correlation.

Seasonal variations in the impacts of the teleconnections could be incorporated by allowing the regression coefficients, $\beta_{3il}(d)$, to vary independently by week or month of year, It seems reasonable to assume that the coefficients are themselves smooth functions of time and this would require yet another smoothing step. Interactions between the teleconnections could also be modeled, but this would further increase the number of parameters making the model even more complex.These extensions clearly provide more flexibility in modeling the relationships between climate and the teleconnection indices, but the number of parameters and computational complexity increases quickly and further statistical developments are required for their implementation.

Modeling the correlation between grid points would provide more precise estimates, and, more importantly, would allow us to compute appropriate measures of uncertainty for quantities obtained from all grid points. In particular, it is necessary to account for spatial correlations in the FPCA. Methods for conducting FPCA with correlated observations have been proposed but are currently too computationally intensive to be applied to such large data sets. Baladandayuthapani et al (2008) developed a Bayesian method implemented via Markov chain Monte Carlo to smooth a set of spatially correlated functions by decomposing each function as a linear combination of B-splines and modeling correlations amongst the coefficients. More recently, Zhou et al (2010) described a classical approach using an EM-algorithm to compute parameter estimates followed by bootstrapping to obtain uncertainty estimates. However, their data contain only 20-40 observations per function and the observations are clustered such that non-zero correlations are restricted to groups of 20-30 functions. In contrast, our data contains 576 observations at each grid point and correlations may extend over the entire province.

Finally, we plan to adapt our methods to study space-time variations in the impact of teleconnections on precipitation. Preciptiation levels are more difficult to model than temperature because precipitation distributions are highly skewed and zero inflated. This is a particular problem in areas like BC because the climate varies considerably. Temperate rainforests on the Pacific coast receive very high levels of precipitation and the regions in the rainshadows of the coastal mountain ranges are very dry. Preliminary analysis of the available precipitation data showed very long periods of drought (many weeks or months) at some stations meaning that the zeros would persist even when the data are aggregated over long periods of time. Alternative solutions include two stage regression models, first modeling the probability of precipitation in a given time period and then modeling the conditional distribution of the precipitation level, as in Williams (1998), or making use of a distribution which simultaneously models both zero inflation and skewness, e.g. the Tweedie distribution used by Dunn (2004).

A Prior Distributions

The following prior distributions were assigned to the parameters of the regression model in equation (2) for each $i = 1,\ldots,168$. The intercept and linear terms in the model of the mean were assigned vague normal priors:

$$\beta_{0i} \sim N(0,30^2) \text{ and } \beta_{1i} \sim N(0,30^2).$$

The same vague prior was also used for the coefficients of the cyclic spline modeling seasonal variations:

$$\beta_{2ij} \sim N(0,30^2), \quad j = 1,\ldots,8.$$

The coefficients governing the impacts of the teleconnection were modeled with a Bayesian P-splines, as described in Section 3.2. The parameter of each curve, $\boldsymbol{b}_{3li}$, $l = 1,\ldots,4$, were assigned the prior structure described in Lang and Brezger (2004):

$$b_{3li}(0) \sim N(0,30^2), \quad b_{3li}(j) \sim N(b_{3li}(j-1),\lambda_{li}^{-2}), \quad j = 1,\ldots,24, \text{ and } \lambda_{li}^2 \sim \Gamma(.05,.005).$$

Parameters of the variance function were also assigned vague normal priors such that:

$$\log(\tau_i^2) \sim N(0,30^2), \quad \rho_{1i} \sim N(0,30^2), \text{ and } \rho_{2ij} \sim N(0,30^2), \quad j = 1,\ldots,8.$$

Acknowledgements S. Bonner and N. Heckmans's research has been supported by the National Science and Engineering Research Council of Canada (NSERC). S. Bonner also received funding from the Pacific Institute for the Mathematical Sciences (PIMS) and the National Science Foundation (NSF Grant No. 0814194), and contract funding from Agriculture and Agri-Food Canada (AAFC). N. Newlands work was supported by AAFC's Sustainable Agricultural Environmental Systems (SAGES), Growing Forward Program. We thank Denise Neilsen, Bill Taylor, Ron Fretwell for the assimilation and quality-control of climate station data.

References

Aguilera A, Escabias M, Valderrama M (2008) Forecasting binary longitudinal data by a functional PC-ARIMA model. Computational Statistics & Data Analysis 52(6):3187–3197

Baladandayuthapani V, Mallick BK, Hong MY, Lupton JR, Turner ND, Carroll RJ (2008) Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. Biometrics 64(March):64–73

Dunn PK (2004) Occurrence and quantity of precipitation can be modelled simultaneously. International Journal of Climatology 24(10):1231–1239

Eilers PHC, Marx BD (1996) Flexible smoothing with B-splines and penalties. Statistical Science 11(2):89–121

Escabias M, Aguilera AM, Valderrama MJ (2005) Modeling environmental data by functional principal component logistic regression. Environmetrics 16(1):95–107

Hancock P, Hutchinson M (2006) Spatial interpolation of large climate data sets using bivariate thin plate smoothing splines. Environmental Modelling & Software 21(12):1684–1694

Haylock M, Hofstra N, Tank A, Klok E, Jones P, New M (2008) A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. Journal of Geophysical Research 113(D20):D20,119

Hutchinson MF (1995) Interpolating mean rainfall using thin plate smoothing splines. International Journal of GIS 9:305–403

Jolliffe I (2002) Principal Component Analysis. Springer-Verlag, New York
Kriss AB, Paul PA, Madden LV (2012) Variability in fusarium head blight epidemics in relation to global climate fluctuations as represented by the El Nino-southern oscillation and other atmospheric patterns. Phytopathology 102(1):55–64
Lang S, Brezger A (2004) Bayesian P-splines. Journal of Computational and Graphical Statistics 13(1):183–212
Ocaña Lara FA, Valderrama MJ, Ocaña Peinado FM, Escabias M (2009) Functional modelling in environmetrics. In: Sakalauskas L, Skiadas C, Zavadskas EK (eds) The XIII International Conference on Applied Stochastic Models and Data Analysis, pp 194–198
Lee D, Shiau J (1994) Thin plate splines with discontinuities and fast algorithms for their computation. SIAM Journal on Scientific Computing 15(6):1311–1330
Little RJA, Rubin DB (2002) Statistical Analysis with Missing Data, 2nd edn. John Wiley, New York
Meiring W (2007) Oscillations and time trends in stratospheric ozone levels. Journal of the American Statistical Association 102(479):788–802
Ramsay JO, Silverman BW (2002) Applied Functional Data Analysis: Methods and Case Studies. Springer-Verlag, New York
Ramsay JO, Silverman BW (2005) Functional Data Analysis, 2nd edn. Springer Series in Statistics, Springer, New York
Ramsay JO, Hooker G, Graves S (2009) Functional Data Analysis with R and MATLAB. UseR!, Springer, New York
Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression. Cambridge University Press, Cambridge, UK Speckman, P, Sun, D
Shabbar A, Khandekar M (1996) The impact of El Nino-southern oscillation on the temperature field over Canada. Atmosphere-Ocean 34(2):401–416
Sheridan S (2003) North American weather-type frequency and teleconnection indices. International Journal of Climatology 23:27–45
Shiau J, Wahba G, Johnson D (1986) Partial spline models for the inclusion of tropopause and frontal boundary information in otherwise smooth two and three dimensional objective analysis. Journal of Atmospheric and Oceanic Technology 3(4):714–725
Temiyasathit C, Kim SB, Park SK (2009) Spatial prediction of ozone concentration profiles. Computational Statistics & Data Analysis 53(11):3892–3906
Valderrama M, FA O, AM A (2002) Forecasting PC-ARIMA models for functional data. In: W H, B R (eds) Proceedings in Computational Statistiscs, Physica-Verlag, pp 25–36
Wang JY, Whitfield PH, Cannon AJ (2006) Influence of Pacific climate patterns on low-flows in British Columbia and Yukon, Canada. Canadian Water Resources Journal 31(1):25–40
Williams PM (1998) Modelling seasonality and trends in daily rainfall data. In: Jordan MI, Kearns MJ, Solla SA (eds) Advances in neural information processing systems Proceedings of the 1997 conference Vol. 10, MIT Press, vol 10, pp 985–991
Wood S (2003) Thin plate regression splines. Journal of the Royal Statistical Society: Series B 65(1):95–114

Wood SN (2006) Generalized Additive Models: An Introduction with R. Texts in Statistical Science, Chapman & Hall/CRC, Boca Raton, Fl

Xu Y, Gao X, Shen Y, Xu C, Shi Y, Giorgi F (2009) A daily temperature dataset over China and its application in validating a RCM simulation. Advances in Atmospheric sciences 26(4):763–772

Zhou L, Huang JZ, Martinez JG, Maity A, Baladandayuthapani V, Carroll RJ (2010) Reduced rank mixed effects models for spatially correlated hierarchical functional data. Journal of the American Statistical Association 105(489):390–400

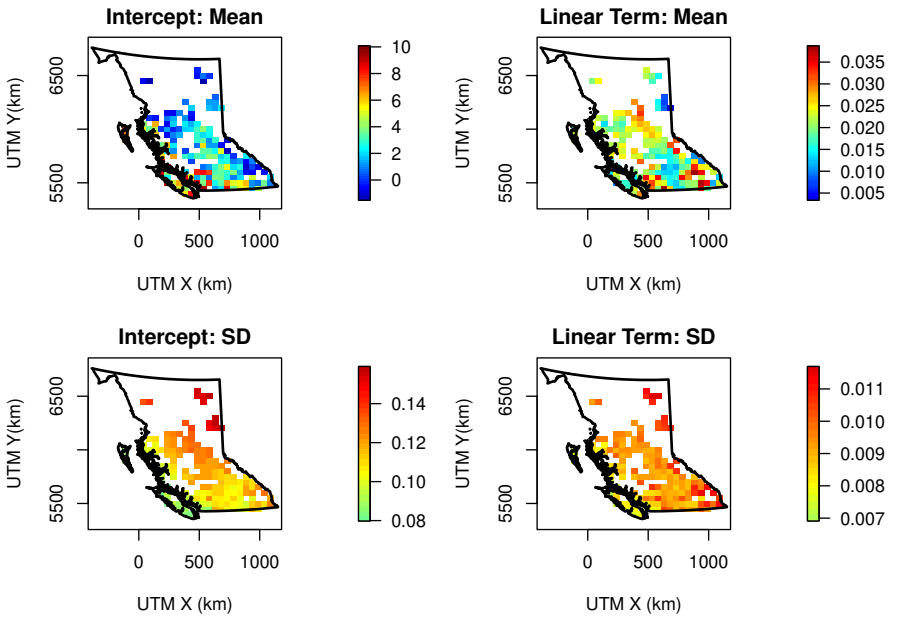

Fig. 1 Estimates and standard errors for the intercept and linear terms obtained from the analysis of the first replicate of the grid data. The left and right panels present the point estimates and standard errors of β_{0i} (top) and β_{1i} (bottom) for each of the 168 grid points obtained from the fitting the separate detrending models and computed using multiple imputation as described in Section 3.3.

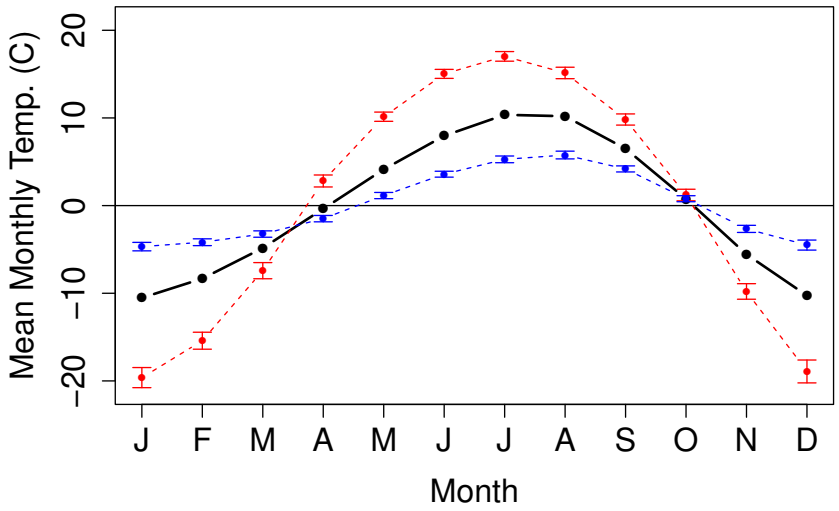

Fig. 2 Plots of the fitted seasonal trends. The black line illustrates the seasonal trend averaged over all of the grid points. The red and blue points represent posterior means of the seasonal trends with 95% CIs at two select grid points as described in the text.

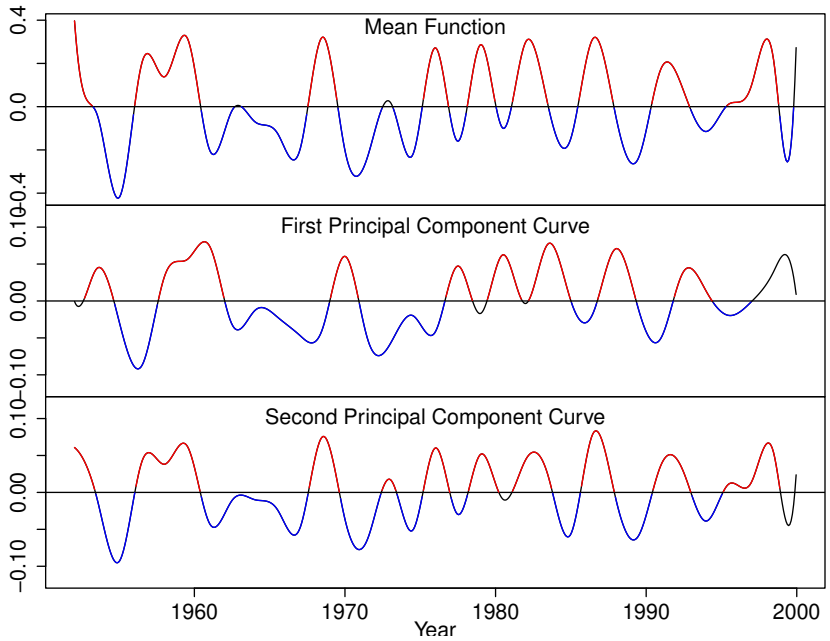

Fig. 3 Plots of the mean function (top panel) and first two principal component curves (bottom panels) for the realized effect of the PNA over the study period. The red and blue segments indicate periods when each curve is above or below 0 °C for more than 12 consecutive months.

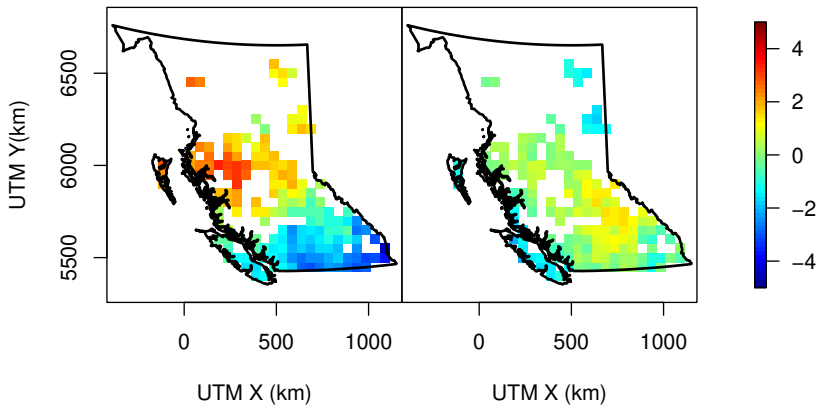

Fig. 4 Scores for the first (left) and second (right) functional principal components of the realized effect of the PNA. The color scale at right is common to both plots.

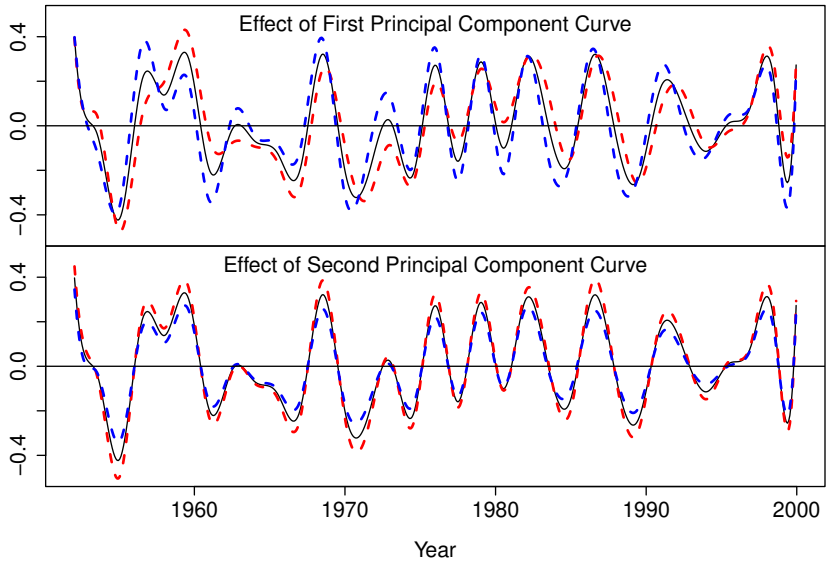

Fig. 5 Impact of the first (top) and second (bottom) principal component curves on the realized effect of the PNA. In each plot the black line represents the mean function. The dashed red/blue lines represent the mean plus/minus the principal component curves times one std. dev. of the corresponding scores.

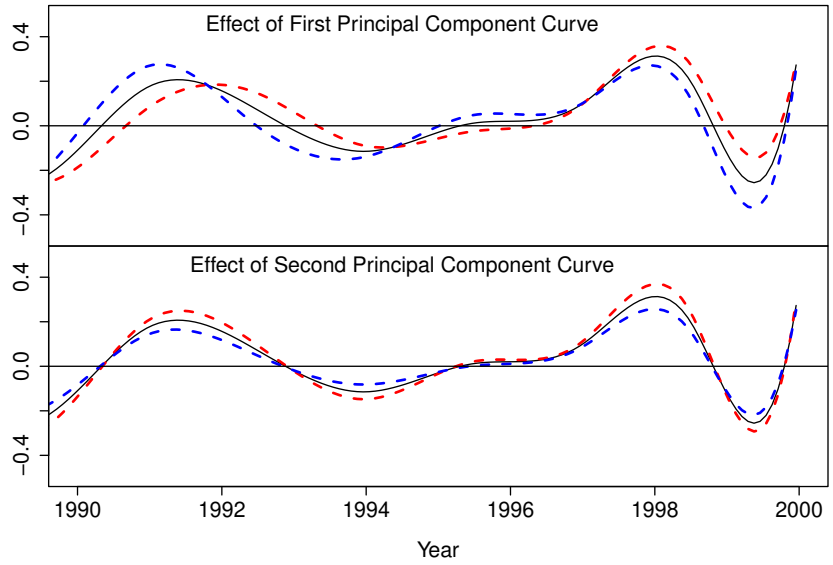

Fig. 6 Impact of the first (top) and second (bottom) principal component curves on the realized effect of the PNA over the last 10 years of the study period. In each plot the black line represents the mean function between 1990 and 2000. The dashed red/blue lines represent the mean plus/minus the principal component curves times one std. dev. of the corresponding scores.