Dimension-Reduced Modeling of Spatio-Temporal Processes

Article *in* Journal of the American Statistical Association · October 2014

DOI: 10.1080/01621459.2014.904232

Jenný Brynjarsdóttir
Case Western Reserve University

Dimension-Reduced Modeling of Spatio-Temporal Processes

Jenný BRYNJARSDÓTTIR and L. Mark BERLINER

The field of spatial and spatio-temporal statistics is increasingly faced with the challenge of very large datasets. The classical approach to spatial and spatio-temporal modeling is very computationally demanding when datasets are large, which has led to interest in methods that use dimension-reduction techniques. In this article, we focus on modeling of two spatio-temporal processes where the primary goal is to predict one process from the other and where datasets for both processes are large. We outline a general dimension-reduced Bayesian hierarchical modeling approach where spatial structures of both processes are modeled in terms of a low number of basis vectors, hence reducing the spatial dimension of the problem. Temporal evolution of the processes and their dependence is then modeled through the coefficients of the basis vectors. We present a new method of obtaining data-dependent basis vectors, which is geared toward the goal of predicting one process from the other. We apply these methods to a statistical downscaling example, where surface temperatures on a coarse grid over Antarctica are downscaled onto a finer grid. Supplementary materials for this article are available online.

KEY WORDS: Bayesian hierarchical modeling; Downscaling; Empirical orthogonal functions; Massive datasets; Maximum covariance patterns; Polar MM5.

1. INTRODUCTION

Challenges associated with the treatment of massive datasets are subjects of intense research. Among these challenges is the development of predictive analyses based on spatio-temporally distributed data. Our focus in this article is the use of one spatio-temporal process to predict another. The challenge of predicting a high-dimensional response from a high-dimensional predictor arises in many disciplines. For example, climate scientists often study interconnected processes using massive datasets derived by remote sensing and computer modeling. A primary subject in neuroscience is the response of the brain, observed by modern scanning methods, to sensory inputs (Ravikumar et al. 2009). Development of models that predict various biological responses to genomic information and other explanatory variables holds promise for understanding a variety of biological systems (Duarte et al. 2007; Richardson, Bottolo, and Rosenthal 2010). Social and behaviorial scientists often seek predictive models of behavior in response to inputs (Lazer et al. 2009). A recent example of parameter calibration in space weather physics (Kleiber et al. 2013) required a predictive model for high-dimensional fields of computer model outputs.

Spatio-temporal analysis has received substantial attention; Cressie and Wikle (2011) provided an authoritative discussion and numerous references. In this article, we add to this literature by exploring and illustrating a class of dimension-reduction methods in combination with Bayesian hierarchical dynamical modeling for space-time prediction.

An abstract statement of the problem is as follows: suppose $\mathcal{Y}(t)$ and $\mathcal{X}(t)$ are spatio-temporally distributed stochastic processes. Conditional on observations of both processes through a time point τ and observations of $\mathcal{X}(t)$ through time $T > \tau$, we wish to provide a predictive distribution for $\mathcal{Y}(t)$ through time T. We assume discrete time and that both stochastic processes are defined on discrete spatial locations. In such cases, linear dynamical spatio-temporal models (DSTM, e.g., Cressie and Wikle 2011) have often been considered. Let $Y_t(s)$ be the value of the $\mathcal{Y}$-process at location s and time t. Set $\mathbf{Y}_t = (Y_t(s_1), \dots, Y_t(s_{N_Y}))'$, where $s_1, \dots, s_{N_Y}$ index N_Y locations where the process is observed. (Though we use the language "spatial" and "location," the formulations can be applied more generally.) A linear DSTM for the $\mathbf{Y}_t$ vectors takes the form

$$\mathbf{Y}_t = M_t \mathbf{Y}_{t-1} + \boldsymbol{\epsilon}_{Y_t}, \quad t = 1, \dots, T, \tag{1}$$

where M_t is an unknown $N_Y \times N_Y$ *transition* (or propagator) matrix and $\boldsymbol{\epsilon}_{Y_t} \overset{\text{iid}}{\sim} N(\mathbf{0}, \Sigma_\epsilon)$. Normality, independence of errors, and constant covariance matrix Σ_ϵ are assumed in this article for simplicity. As the dimension N_Y grows, effective modeling and estimation of M_t and Σ_ϵ become very difficult and computationally demanding.

Covariate information can be incorporated in (1) in a variety of ways. For example, M_t may be parameterized to depend on the $\mathcal{X}$-process; functions of the $\mathcal{X}$-process can be included directly into the mean of $\mathbf{Y}_t$; the distribution of the errors $\boldsymbol{\epsilon}_{Y_t}$ may be modeled to depend on the $\mathcal{X}$-process; and various combinations of such notions. In any case, increasing dimensions of both the $\mathcal{Y}$- and $\mathcal{X}$-processes conspire to make the problem difficult.

A class of dimension-reduced alternatives to the DSTM in (1) involves two stages. First, each $\mathbf{Y}_t$ is assumed to have a representation of the form

$$\mathbf{Y}_t = U_t \mathbf{a}_t + \boldsymbol{\eta}_{Y t}, \quad t = 1, \dots, \tau, \tag{2}$$

Jenný Brynjarsdóttir is Assistant Professor, Department of Mathematics, Applied Mathematics and Statistics, Case Western Reserve University, Cleveland, OH 44106 (E-mail: *jenny.brynjarsdottir@case.edu*). Mark Berliner is Professor and Chair, Department of Statistics, The Ohio State University, Columbus, OH 43210 (E-mail: *mb@stat.osu.edu*). This research was supported by the National Science Foundation under grants ATM-07-24403 and DMS-10-49064. The authors thank Peter Craigmile, Noel Cressie, and Steve MacEachern for valuable input during various stages of the development of this work, and the editor, associate editor, and two anonymous reviewers for insightful and helpful comments.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.

Journal of the American Statistical Association
December 2014, Vol. 109, No. 508, Theory and Methods
DOI: 10.1080/01621459.2014.904232

where U_t is an $N_Y \times K_Y$-dimensional matrix, $\mathbf{a}_t$ is a K_Y-dimensional random vector with $K_Y \ll N_Y$, and $\boldsymbol{\eta}_{Y_t}$ is a zero-mean error vector. Second, an accompanying model for $\{\mathbf{a}_1, \ldots, \mathbf{a}_T\}$ is developed. The key is that the U_t are specified (see Section 1.1) so much of the critical modeling and analysis is transferred from the large dimension N_Y to the far smaller dimension K_Y.

Our goal of incorporating the $\mathcal{X}$-process into predictions of the $\mathcal{Y}$-process is typically complicated when the $\mathcal{X}$-process is also high dimensional. For $t = 1, \ldots, T$, let $\mathbf{X}_t$ be an N_X-dimensional covariate vector. Our suggestion involves two basic steps. First, form a dimension-reduced model analogous to (2):

$$\mathbf{X}_t = V_t \mathbf{b}_t + \boldsymbol{\eta}_{Xt}, \quad t = 1, \ldots, \tau, \tag{3}$$

where V_t is an $N_X \times K_X$-dimensional matrix and $\mathbf{b}_t$ is a K_X-dimensional random vector, where $K_X << N_X$. The second step is to develop a model for the coefficients $\mathbf{a}_t$ in (2) conditioned on $\mathbf{b}_t$. That is, our incorporation of covariate information takes place in K_Y and K_X dimensions rather than in N_Y and N_X dimensions. Our challenges are to develop effective and efficient approaches to choosing both dimension reductions (i.e., the matrices U_t; and V_t in (2) and (3)).

Dimension-reduction techniques differ mostly in the choice of U_t matrices (see Equation (2)). Examples include principal components (PCs; Wikle and Cressie 1999; Berliner, Wikle, and Cressie 2000), locally weighted mixtures of linear regressions (Stroud, Müller, and Sansó 2001), multiresolution (wavelet) bases (Wikle et al. 2001; Kang, Cressie, and Shi 2010), and Fourier basis functions (Xu, Wikle, and Fox 2005). The process convolution methods of Ver Hoef and Barry (1998), Higdon (1998), and Calder (2007) can also be viewed as dimension-reduction approaches to the DSTM. Wikle (2010) gave a more extensive overview of dimension-reduced spatial modeling and Cressie and Wikle (2011) gave a detailed discussion of dimension-reduced spatio-temporal modeling. Finally, we note that the model defined in (2) is related to the spatio-temporal random effects model proposed by Cressie, Shi, and Kang (2010) for which efficient filtering and smoothing methods have been developed (Kang, Cressie, and Shi 2010; Katzfuss and Cressie 2011, 2012).

A Bayesian hierarchical modeling framework coupled with dimension reduction is presented in Section 2. Our suggestions for selection of the dimension reductions are motivated by the criterion of mean squared prediction error. That key discussion is in Section 2.1. We indicate how selected multivariate methods such as PCs and canonical correlation analysis (CCA) can be used for dimension reduction of space-time processes in Section 3. In Sections 3.2 and 3.3, we present our suggestions for dimension reduction for prediction. In Section 4, we apply our methods in a *statistical downscaling* example; specifically using large-scale global information (our X-process) to predict surface temperatures over the Antarctic at a fine-scale (our Y-process). Section 5 contains discussion.

2. HIERARCHICAL MODELING FRAMEWORK

We outline a hierarchical modeling framework that incorporates dimension reduction. For presentation purposes, we first consider two spatial-only processes and briefly discuss adding the temporal dimension at the end of this section. In Section 2.1, we discuss mean squared error (MSE) for prediction using the dimension-reduction approach.

First, we establish some notation. Let $Y(s)$ and $X(c)$ be spatial processes at locations s and c. Let $\mathbf{Y}$ and $\mathbf{X}$ be random vectors containing values of the processes at N_Y and N_X locations, that is, $Y_i = Y(s_i)$, for $i = 1, \ldots, N_Y$ and $X_j = X(c_j)$, for $j = 1, \ldots, N_X$. The notation $\left[Y|X\right]$ stands for "the conditional probability density function of Y given X" and we use U' to denote the matrix transpose of U.

A general Bayesian probability model for $\mathbf{Y}$, $\mathbf{X}$, and a vector of unknown parameters $\boldsymbol{\theta}$ can be written as $\left[\mathbf{Y}, \mathbf{X}|\boldsymbol{\theta}\right]\left[\boldsymbol{\theta}\right]$. We can model the likelihood hierarchically since

$$\left[\mathbf{Y}, \mathbf{X}|\boldsymbol{\theta}\right] = \left[\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}\right]\left[\mathbf{X}|\boldsymbol{\theta}\right]. \tag{4}$$

Modeling each of $[\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}]$ and $[\mathbf{X}|\boldsymbol{\theta}]$ eliminates the need to specify a joint model for $\mathbf{Y}$ and $\mathbf{X}$ directly, which can be especially challenging for multivariate spatial processes (see, e.g., Royle and Berliner 1999; Cressie and Wikle 2011). This approach is also very natural when scientific knowledge exists for how $\mathbf{Y}$ depends on $\mathbf{X}$ (see, e.g., Berliner 2003). However, even with the hierarchical approach, fitting a spatial model for $\left[\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}\right]$ with high-dimensional $\mathbf{Y}$ and $\mathbf{X}$ is difficult. For example, if we consider a normal linear model,

$$\mathbf{Y} = F\mathbf{X} + \boldsymbol{\varepsilon}, \quad \text{where } \boldsymbol{\varepsilon} \sim N_{N_Y}(\mathbf{0}, \Sigma_\varepsilon), \tag{5}$$

we need to estimate an $N_Y \times N_X$-dimensional model matrix F and the $N_Y \times N_Y$-dimensional covariance matrix Σ_ε. In high dimensions, F and Σ_ε are typically formulated to depend on a comparatively low-dimensional parameter vector. Here, we explore the alternative of a dimension-reduction approach for both $\mathbf{Y}$ and $\mathbf{X}$.

Let $\mathcal{U}$ and $\mathcal{V}$ be $N_Y \times N_Y$- and $N_X \times N_X$-dimensional matrices that contain basis vectors that span $\mathbb{R}^{N_Y}$ and $\mathbb{R}^{N_X}$, respectively, and let the matrices U and V contain selections of K_Y and K_X column vectors of $\mathcal{U}$ and $\mathcal{V}$. Dimension-reduction models are defined as

$$\mathbf{Y} = \sum_{k=1}^{K_Y} \mathbf{u}_k a_k + \boldsymbol{\eta}_Y = U\mathbf{a} + \boldsymbol{\eta}_Y \quad \text{and}$$
$$\mathbf{X} = \sum_{k=1}^{K_X} \mathbf{v}_k b_k + \boldsymbol{\eta}_X = V\mathbf{b} + \boldsymbol{\eta}_X, \tag{6}$$

where $\mathbf{a}$ and $\mathbf{b}$ are K_Y- and K_X-dimensional unknown vectors of coefficients, often called *amplitudes*, and the $\boldsymbol{\eta}$'s are random vectors modeled to account for the left over structure. The dimension reductions are often judged to be effective if the K_Y and K_X basis vectors in U and V capture the most essential structure of $\mathbf{Y}$ and $\mathbf{X}$, respectively. However, we also are concerned with predictive power. In Section 3, we discuss some options for choosing U and V.

The dimension-reduced Bayesian hierarchical framework in this article takes the form

$$[\mathbf{Y}, \mathbf{X}, \mathbf{a}, \mathbf{b}, \boldsymbol{\theta}] = [\mathbf{Y}, \mathbf{X}|\mathbf{a}, \mathbf{b}, \boldsymbol{\theta}][\mathbf{a}, \mathbf{b}|\boldsymbol{\theta}][\boldsymbol{\theta}] \tag{7}$$
$$\overset{A}{=} [\mathbf{Y}|\mathbf{a}, \boldsymbol{\theta}][\mathbf{X}|\mathbf{b}, \boldsymbol{\theta}][\mathbf{a}|\mathbf{b}, \boldsymbol{\theta}][\mathbf{b}|\boldsymbol{\theta}][\boldsymbol{\theta}]. \tag{8}$$

The equality in (7) is a simple fact; we arrive at (8) based on assumptions explained below. We refer to the three terms on the right-hand side of (7) as the *data model*, *process model*, and *parameter model*, respectively (Berliner 1996).

As indicated in (8), the process model, $\big[\mathbf{a}, \mathbf{b}|\boldsymbol{\theta}\big]$, can be developed hierarchically. Continuing with the linear model in (5), we note that the dimension reduction in (6) implies a linear model for $\big[\mathbf{a}|\mathbf{b}, \boldsymbol{\theta}\big]$. Inserting the expressions for $\mathbf{Y}$ and $\mathbf{X}$ (6) into the model in (5), we get

$$U\mathbf{a} + \boldsymbol{\eta}_Y = FV\mathbf{b} + F\boldsymbol{\eta}_X + \boldsymbol{\varepsilon}. \tag{9}$$

$$\Rightarrow\ \mathbf{a} = U'FV\mathbf{b} + U'F\boldsymbol{\eta}_X + U'\boldsymbol{\varepsilon} - U'\boldsymbol{\eta}_Y, \tag{10}$$

if the column vectors in U are orthonormal. Defining $H = U'FV$ and pooling the remaining error terms, we obtain a linear model for the amplitudes:

$$\mathbf{a} = H\mathbf{b} + \mathbf{e}. \tag{11}$$

The model for vectors of amplitudes in (11) crystallizes the operational benefits of using dimension reduction for $\mathbf{Y}$ and $\mathbf{X}$ if both K_Y and K_X are very small compared to N_Y and N_X. The difficulties lie in choosing basis vectors U and V that are simultaneously low dimensional and warrant the conditional independence assumptions made in the data model part of (8), namely, that

$$\big[\mathbf{Y}, \mathbf{X}|\mathbf{a}, \mathbf{b}, \boldsymbol{\theta}\big] = \big[\mathbf{Y}|\mathbf{a}, \boldsymbol{\theta}\big]\big[\mathbf{X}|\mathbf{b}, \boldsymbol{\theta}\big]. \tag{12}$$

This is perhaps the most critical assumption underlying our approach. To clarify the interaction between the dimension reductions and conditional independence, we consider a more common notion. Rather than (7) and (8), consider

$$[\mathbf{Y}, \mathbf{X}, \boldsymbol{\mu}_Y, \boldsymbol{\mu}_X, \boldsymbol{\theta}] = [\mathbf{Y}, \mathbf{X}|\boldsymbol{\mu}_Y, \boldsymbol{\mu}_X, \boldsymbol{\theta}][\boldsymbol{\mu}_Y, \boldsymbol{\mu}_X|\boldsymbol{\theta}][\boldsymbol{\theta}] \tag{13}$$

$$\begin{aligned} &\overset{A}{=} [\mathbf{Y}|\boldsymbol{\mu}_Y, \boldsymbol{\theta}][\mathbf{X}|\boldsymbol{\mu}_X, , \boldsymbol{\theta}][\boldsymbol{\mu}_Y|\boldsymbol{\mu}_X, \boldsymbol{\theta}] \\ &\quad \times [\boldsymbol{\mu}_X|\boldsymbol{\theta}][\boldsymbol{\theta}], \end{aligned} \tag{14}$$

where $\boldsymbol{\mu}_Y$ and $\boldsymbol{\mu}_X$ are the means of $\mathbf{Y}$ and $\mathbf{X}$, respectively. Suppose our model is that $\mathbf{Y}$ and $\mathbf{X}$ are conditionally independent, noisy observations of their means and that those means are related by the model

$$\boldsymbol{\mu}_Y = F\boldsymbol{\mu}_X + \boldsymbol{\varepsilon}, \quad \text{where } \boldsymbol{\varepsilon} \sim N_{N_Y}(\mathbf{0}, \Sigma_\varepsilon) \tag{15}$$

(recall (5)). The conditional independence assumptions leading to (14) are quite natural: given $\boldsymbol{\mu}_X$, $\mathbf{X}$ provides no additional information about $\mathbf{Y}$ and similarly, due to (15), given $\boldsymbol{\mu}_Y$, $\boldsymbol{\mu}_X$ provides no information additional information about $\mathbf{Y}$. The issue here and in similar applications of dimension reduction is whether or not such natural assumptions are plausible after the reductions. That is, do the losses in statistical information associated with our reductions lead to implied error vectors with structural relationships that yield significant dependences? Intuitively, such dependences may be minor if the dimension-reduction models in (6) are a good representation of the structures in $\mathbf{Y}$ and $\mathbf{X}$. We offer an assessment of the conditional independence assumptions in Section 4. That check leads us to conclude that the departures from conditional independence in our example are negligible.

Incorporating the temporal dimension is straightforward but introduces even more interesting modeling issues. Essentially, we add temporal dependence in all the models used in the purely spatial case. We consider only discrete time and for ease of notation, we assume that Y- and X-processes are observed at the same time points $t = 1, \ldots, T$. Let $\mathbf{Y}_t$ and $\mathbf{X}_t$ be N_{Y_t}- and N_{X_t}-dimensional vectors of observations at those times, respectively. Let U_t and V_t be matrices of K_{Y_t} and K_{X_t}-selected basis vectors for the Y- and X-processes, respectively. The dimension-reduced models are

$$\mathbf{Y}_t = \sum_{k=1}^{K_{Y_t}} \mathbf{u}_{t,k} a_{t,k} + \boldsymbol{\eta}_{Y_t} = U_t\mathbf{a}_t + \boldsymbol{\eta}_{Y_t} \tag{16}$$

$t = 1, \ldots, T$ and

$$\mathbf{X}_t = \sum_{k=1}^{K_{Xt}} \mathbf{v}_{t,k} b_{t,k} + \boldsymbol{\eta}_{X_t} = V_t\mathbf{b}_t + \boldsymbol{\eta}_{X_t} \tag{17}$$

$t = 1, \ldots, T$. Although we allow for time-dependent basis vectors U_t and V_t, we may want to use the same basis vectors for some or all time point in practice. The temporal dependence introduces the need for more assumptions of conditional independence. First, our framework assumes that for each t, $\mathbf{Y}_t$ and $\mathbf{X}_t$ are conditionally mutually independent, given all amplitude vectors, that is, data model is

$$\begin{aligned} &[\mathbf{Y}_1, \ldots, \mathbf{Y}_T, \mathbf{X}_T, \ldots, \mathbf{X}_T|\mathbf{a}_1, \ldots \mathbf{a}_T, \mathbf{b}_1, \ldots \mathbf{b}_T, \boldsymbol{\theta}] \\ &\quad = \prod_{t=1}^{T} [\mathbf{Y}_t|\mathbf{a}_t, \boldsymbol{\theta}]\ [\mathbf{X}_t|\mathbf{b}_t, \boldsymbol{\theta}]\ . \end{aligned} \tag{18}$$

The individual distributions represent the dimension-reduction models in (16) and (17). Second, the $\mathbf{a}_t$ and $\mathbf{b}_t$ vectors can be modeled in a variety of ways. Autocorrelated temporal dependence with a small lag, for example, a first-order Markovian model for the pairs $(\mathbf{a}_t', \mathbf{b}_t')'$, is a convenient choice. Another useful notion is to let the sequence of $\mathbf{b}_t$ carry the temporal dependence (conditional on $\boldsymbol{\theta}$). For example, in a first-order Markovian model, assume for each t that

$$[\mathbf{a}_t, \mathbf{b}_t|\mathbf{a}_{t-1}, \mathbf{b}_{t-1}, \boldsymbol{\theta}] = [\mathbf{a}_t|\mathbf{b}_t, \boldsymbol{\theta}]\ [\mathbf{b}_t|\mathbf{b}_{t-1}, \boldsymbol{\theta}]\ . \tag{19}$$

2.1 Predictive Mean Squared Error

Recall that the best predictor, in terms of MSE, of a random vector $\mathbf{Y}$ based on a random vector $\mathbf{X}$ is $E(\mathbf{Y}|\mathbf{X})$ (e.g., Rao 1973, sec. 4g.1). Under the linear model in (5), the optimal predictor is $E\big(\mathbf{Y}|\mathbf{X}\big) = F\mathbf{X}$. Under the hierarchical model (8), we find

$$\begin{aligned} E(\mathbf{Y}|\mathbf{X}) &= E(E(\mathbf{Y}|\mathbf{a}, \mathbf{b}, \boldsymbol{\theta}, \mathbf{X})|\mathbf{X}) && (20) \\ &= E(E(U\mathbf{a}|\mathbf{b}, \boldsymbol{\theta})|\mathbf{X}) && (21) \\ &= UE(H\mathbf{b}|\mathbf{X}); && (22) \end{aligned}$$

recall that H may contain unknown parameters included in $\boldsymbol{\theta}$ and therefore remains inside the final expectation.

Insight into desirable properties of the models can be gained by considering the MSE under the Bayesian model (8):

$$\begin{aligned} &E(||\mathbf{Y} - E(\mathbf{Y}|\mathbf{X})||^2) \\ &\quad = E(||U\mathbf{a} - UE(H\mathbf{b}|\mathbf{X}) + \boldsymbol{\eta}_Y||^2)\ \text{(see (6))} && (23) \\ &\quad = E(||U(\mathbf{a} - E(H\mathbf{b}|\mathbf{X}))||^2) + E(||\boldsymbol{\eta}_Y||^2). && (24) \end{aligned}$$

This expression highlights the role of the quality of the dimension reduction for $\mathbf{Y}$, that is, small variance of η_{Yi}, as well as the quality of predictions of the amplitudes $\mathbf{a}$ based on $\mathbf{X}$, but through the amplitudes $\mathbf{b}$. Intuitively, the latter point suggests that strong correlations between the amplitude vectors $\mathbf{a}$ and $\mathbf{b}$ are desirable.

For additional insight, note that the form of the Bayesian predictor suggests a general class of predictive procedures. Under the models $\mathbf{Y} = F\mathbf{X} + \boldsymbol{\varepsilon}$ and $\mathbf{X} = V\mathbf{b} + \boldsymbol{\eta}_X$, consider an estimator of $\mathbf{b}$, say $\widehat{\mathbf{b}}(\mathbf{X}) = V'\mathbf{X}$ and some estimator $\hat{H}$ of H. If the columns of V are orthonormal, a simple predictor of $\mathbf{Y}$ is $\hat{\mathbf{Y}} = U\hat{H}V'\mathbf{X}$. The *increase* in MSE from using this predictor rather than the optimal rule is

$$\begin{aligned} &E(||E(\mathbf{Y}|\mathbf{X}) - U\hat{H}V'\mathbf{X}||^2) \\ &\quad = E[\mathbf{X}'(F - U\hat{H}V')'(F - U\hat{H}V')\mathbf{X}] \qquad (25) \\ &\quad = E[\mathbf{X}'(F - UHV' + U(H-\hat{H})V')' \\ &\qquad \times (F - UHV' + U(H-\hat{H})V')\mathbf{X}]. \qquad (26) \end{aligned}$$

The terms in (26) are another indicator of the relevance of both the quality of the dimension-reduction approximation (i.e., $F - UHV'$) and the estimation error of H (i.e., $H - \hat{H}$). In typical cases, we expect these two features to compete; dimension reductions based on many terms may improve the approximation but increase the relative difficulty in modeling and estimation of H.

3. DATA-DEPENDENT BASIS VECTORS

Here, we turn to the issue of selecting basis vectors U_t and V_t for the dimension-reduction models in (16) and (17). We take the approach of *data-dependent* basis vectors, that is, patterns that are obtained from existing data. In Section 3.1, we give a brief discussion of the use of PCs, maximum covariance analysis (MCA), and canonical covariance analysis. In Sections 3.2 and 3.3, we develop *maximum covariance patterns* (MCPs), a new approach to obtaining data-dependent basis vectors.

3.1 Common Dimension-Reduction Techniques

The most familiar dimension-reduction technique is perhaps the method of PCs, also known as empirical orthogonal functions (EOFs). EOFs are used in many disciplines, but particularly in the climatic science (e.g., Zhang, Wallace, and Battisti 1997; Esteban-Parra, Rodrigo, and Castro-Diez 1998). EOFs are the eigenvectors of the empirical covariance matrix, S_{YY}, of a random vector $\mathbf{Y}$. The usual approach is to treat data observed at different time points as *repeated measurements* (i.e., act as if the $\mathbf{Y}_t$ are temporally stationary). Note that storage of the full $N_Y \times N_Y$ sample covariance matrix is not necessary, to obtain the eigenvectors one can apply *singular value decomposition* (Golub and Van Loan 1996) directly on the $N_Y \times T$-dimensional data matrix. The first few EOFs are candidates for columns of U in the dimension-reduced modeling of the $\mathbf{Y}_t$. Similarly, the EOFs for the $\mathbf{X}$-process data are candidates for the dimension-reduced modeling of the $\mathbf{X}_t$. It is important to keep in mind that the proportion of sample variance explained is an average over the T time points. One approach to adapting EOF analysis to deal with potentially nonstationary processes involves approximating the mean of $\mathbf{Y}_t$ as a linear combination of a selection of EOFs where the coefficients are time varying (e.g., Berliner, Wikle, and Cressie 2000; Cressie and Wikle 2011). This is how we proceed in this article.

MCA and CCA are popular dimension-reduction tools for studying the relationship between two random vectors. For the datasets $\mathbf{Y}_t\,; t = 1, \ldots, T$ and $\mathbf{X}_t\,; t = 1, \ldots, T$, let Y (X) be the $N_Y \times T$ $(N_X \times T)$ temporally centered data matrix. Also, define the sample cross-covariance matrix $S_{YX} = \frac{1}{T-1}YX'$. In MCA, the left and right singular vectors obtained from the singular value decomposition of S_{YX} are pairs of *MC patterns*. They are interpreted as follows: first, find a pair of linear combinations $\mathbf{u}_1'\mathbf{Y}$ and $\mathbf{v}_1'\mathbf{X}$ that maximizes their covariance over all such linear combinations. Next, we find a pair $\mathbf{u}_2'\mathbf{Y}$ and $\mathbf{v}_2'\mathbf{X}$ that maximizes their covariance over all such linear combinations but subject to the conditions that $\mathbf{u}_1$ and $\mathbf{u}_2$ are orthogonal and $\mathbf{v}_1$ and $\mathbf{v}_2$ are orthogonal. The procedure continues iteratively. For further discussion overview and references see, von Storch and Zwiers (2001), Johnson and Wickern (2002), and Cressie and Wikle (2011).

The first few pairs of $\mathbf{u}$ and $\mathbf{v}$ vectors (obtained either by MCA or CCA) are candidates for basis vectors for the model in (16) and (17). Since they lead to models in which the covariances (or correlations) between the $\mathbf{a}$ and $\mathbf{b}$ amplitude vectors are high, they seem to accomplish one of the desirable features discussed in the second to last paragraph of Section 2.1. However, it is not clear that these basis vectors offer useful dimension-reduced approximations (recall the last paragraph of Section 2.1).

3.2 Maximum Covariance Patterns

We now develop data-dependent basis vectors that are tailored to the objective of predicting the $\mathbf{Y}$ process from observations of the $\mathbf{X}$ process. The suggestion is based on the heuristics presented in Section 2.1. Our approach is described in two steps:

1. Develop a dimension reduction for $\mathbf{Y}$. We suppose that a collection of normalized patterns $\mathbf{u}_1, \ldots, \mathbf{u}_{K_Y}$ has been specified. For example, $\mathbf{u}_1, \ldots, \mathbf{u}_{K_Y}$ could be the K_Y leading EOFs obtained from the data matrix Y as described in Section 3.1.
2. For each vector $\mathbf{u}_k$ find the normalized vector $\mathbf{v}_k$ that maximizes the covariance of the amplitudes, that is, we maximize $\mathrm{cov}(a_k, b_k) = \mathrm{cov}(\mathbf{u}_k'\mathbf{Y}, \mathbf{v}_k'\mathbf{X})$.

The solution is stated in the following theorem.

Theorem 3.1. Let $\mathbf{Y}$ and $\mathbf{X}$ be N_Y- and N_X-dimensional random vectors with cross-covariance matrix $\mathrm{cov}(\mathbf{Y}, \mathbf{X}) = \Sigma_{YX}$. Let $\mathbf{u}$ be a fixed, normalized N_Y-dimensional vector. The linear combination $\mathbf{v}'\mathbf{X}$ that maximizes the covariance $\mathrm{cov}(\mathbf{u}'\mathbf{Y}, \mathbf{v}'\mathbf{X})$, under the constraint that $\mathbf{v}$ is of length 1, is

$$\mathbf{v}^* = \frac{\Sigma_{XY}\mathbf{u}}{||\Sigma_{XY}\mathbf{u}||}\,, \qquad (27)$$

where $||\cdot||$ is the Euclidean norm and $\Sigma_{XY} = \Sigma_{YX}'$.

Proof. Since $\mathbf{c}'\mathbf{v} = ||\mathbf{c}||\,||\mathbf{v}||\cos(\angle(\mathbf{c}, \mathbf{v}))$ it follows that for fixed $\mathbf{c}$ the vector $\mathbf{v}^* = \frac{\mathbf{c}}{||\mathbf{c}||}$ maximizes $\mathbf{c}'\mathbf{v}$ under the constraint that $||\mathbf{v}|| = 1$. Then (27) follows directly, since $\mathrm{cov}(\mathbf{u}'\mathbf{Y}, \mathbf{v}'\mathbf{X}) = (\Sigma_{XY}\mathbf{u})'\,\mathbf{v}$. □

We apply this result using the empirical cross-covariance matrix $S_{YX} = \frac{1}{T-1}YX'$, where Y and X are the $N_Y \times T$ and $N_X \times T$, temporally centered, data matrices.

Definition 3.2. Given a collection of normalized vectors $\mathbf{u}_1, \ldots, \mathbf{u}_K$, we define MCPs $\mathbf{v}_1, \ldots, \mathbf{v}_K$ as

$$\mathbf{v}_k = \frac{S_{XY}\mathbf{u}_k}{||S_{XY}\mathbf{u}_k||} \quad k = 1, \ldots, K. \tag{28}$$

We add the constraints that the $\mathbf{v}$ are mutually orthogonal in Section 3.3.

3.2.1 Proportion of Total Variance Explained. The motivation behind MCPs was our interest in predicting $\mathbf{Y}$. We are not necessarily concerned with how much of the total sample variation of $\mathbf{X}$ is explained by the MCPs. Rather, we want to answer the question "What proportion of the total variance in $\mathbf{Y}$ is explained by the $\mathbf{v}_K$ vectors through the $\mathbf{u}_K$ vectors?" With that in mind we introduce the following definition:

Definition 3.3. Let $\mathbf{Y}$ and $\mathbf{X}$ be N_Y and N_X-dimensional random vectors and let $\mathbf{u}$ and $\mathbf{v}$ be specified N_Y and N_X-dimensional vectors. We define the *proportion of the total variance in* $\mathbf{Y}$ *that is explained by* $\mathbf{v}$ *through* $\mathbf{u}$ as

$$p_{\mathbf{u}}(\mathbf{v}) = \frac{\text{cov}(\mathbf{u}'\mathbf{Y}, \mathbf{v}'\mathbf{X})^2}{\text{var}(\mathbf{v}'\mathbf{X})\,\text{tr}(\Sigma_{YY})}. \tag{29}$$

The proportion $p_{\mathbf{u}}(\mathbf{v})$ is the total variance of $\mathbf{Y}$ that is explained by $\mathbf{u}$ $(\text{var}(\mathbf{u}'\mathbf{Y})/\text{tr}(\Sigma_{YY}))$ times the squared correlation between amplitudes $A = \mathbf{u}'\mathbf{Y}$ and $B = \mathbf{v}'\mathbf{X}$. This definition leads to the following result.

Theorem 3.4. Let $\mathbf{Y}$ and $\mathbf{X}$ be N_Y- and N_X-dimensional random vectors and let $\mathbf{u}$ be a specified N_Y-dimensional vector. Let $\mathbf{v}$ be the MCP of $\mathbf{X}$ with respect to $\mathbf{u}$ and let $\mathbf{e}$ be the first EOF of $\mathbf{X}$. Then

$$p_{\mathbf{u}}(\mathbf{v}) \geq p_{\mathbf{u}}(\mathbf{e}). \tag{30}$$

Proof. Since $\text{var}(\mathbf{v}'\mathbf{X}) \leq \text{var}(\mathbf{e}'\mathbf{X})$ and $\text{cov}(\mathbf{u}'\mathbf{Y}, \mathbf{v}'\mathbf{X}) \geq \text{cov}(\mathbf{u}'\mathbf{Y}, \mathbf{e}'\mathbf{X})$, we have

$$p_{\mathbf{u}}(\mathbf{v}) = \frac{\text{cov}(\mathbf{u}'\mathbf{Y}, \mathbf{v}'\mathbf{X})^2}{\text{var}(\mathbf{v}'\mathbf{X})\,\text{tr}(\Sigma_{YY})} \geq \frac{\text{cov}(\mathbf{u}'\mathbf{Y}, \mathbf{e}'\mathbf{X})^2}{\text{var}(\mathbf{e}'\mathbf{X})\,\text{tr}(\Sigma_{YY})} = p_{\mathbf{u}}(\mathbf{e}). \tag{31}$$

Theorem 3.4 shows that in terms of total variance in $\mathbf{Y}$ explained, it is better to use the MCP as the first basis vector for $\mathbf{X}$ rather than the first EOF. □

3.3 Orthogonal Maximum Covariance Patterns

Orthogonal basis functions yield advantages in interpretation and computation.

Definition 3.5. Given a collection of normalized vectors $\mathbf{u}_1, \ldots, \mathbf{u}_K$, define *orthogonal maximum covariance patterns (OMCPs)* $\mathbf{v}_1^o, \ldots, \mathbf{v}_K^o$, where $\mathbf{v}_1^o$ is the first MCP $\mathbf{v}_1$ and

$$\mathbf{v}_k^o = \frac{1}{L_k}\left(S_{XY}\mathbf{u}_k - \sum_{j=1}^{k-1}((S_{XY}\mathbf{u}_k)'\mathbf{v}_j)\mathbf{v}_j\right), \tag{32}$$

$k = 2, \ldots, K$, where L_k is the normalizing constant.

Notice that the OMCPs are in fact the Gram–Schmidt orthogonalization of the MCPs. The following result provides justification of the recipe for obtaining the OMCPs.

Theorem 3.6. Let $\mathbf{Y}$ and $\mathbf{X}$ be N_Y- and N_X-dimensional random vectors with cross-covariance matrix $\text{cov}(\mathbf{Y}, \mathbf{X}) = \Sigma_{YX}$. Let $\mathbf{u}_1, \ldots, \mathbf{u}_K$ be fixed, linearly independent N_Y-dimensional vectors. Set $\mathbf{v}_1$ to be the MCP corresponding to $\mathbf{u}_1$. For each $k = 2, \ldots, K$, the unit-length vector $\mathbf{v}_k$ that maximizes the covariance $\text{cov}(\mathbf{u}_k'\mathbf{Y}, \mathbf{v}_k'\mathbf{X})$ under the constraint that $\mathbf{v}_k$ is orthogonal to $\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}$ is

$$\mathbf{v}_k = \frac{1}{L_k}\left(\Sigma_{XY}\mathbf{u}_k - \sum_{j=1}^{k-1}((\Sigma_{XY}\mathbf{u}_k)'\mathbf{v}_j)\mathbf{v}_j\right), \tag{33}$$

where $\Sigma_{XY} = \Sigma_{YX}'$ and L_k is the normalizing constant.

Proof. Note again that $\mathbf{c}'\mathbf{v}_K = ||\mathbf{c}||\,||\mathbf{v}_K||\cos(\angle(\mathbf{c}, \mathbf{v}_K))$. The optimal vector is both orthogonal to $\mathbf{v}_j$, $j = 1, \ldots, K-1$ and has the smallest $\angle(\mathbf{c}, \mathbf{v}_K)$, making $\mathbf{c}'\mathbf{v}_K$ as big as possible. The solution is therefore the residual vector from the orthogonal projection of $\mathbf{c}$ onto the space spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_{K-1}$. Substitute $\mathbf{c}$ with $\Sigma_{XY}\mathbf{u}_k$ and we have the result in (33). □

The constraint of orthogonality lowers the covariance of the amplitudes compared with the covariance of MCP amplitudes. Let $\mathbf{v}_k^*$ be the MCP of $\mathbf{X}$ with respect to $\mathbf{u}_k$ and let $\mathbf{v}_k^o$ be the OMCP. Recalling that $||\mathbf{v}_k^*|| = ||\mathbf{v}_k^o|| = 1$, we see that

$$\begin{aligned}\text{cov}(\mathbf{u}_k'\mathbf{Y}, \mathbf{v}_k^{o\prime}\mathbf{X}) &= ||\Sigma_{XY}\mathbf{u}_k||\cos(\angle(\Sigma_{XY}\mathbf{u}_k, \mathbf{v}_k^o)) \\ &= \text{cov}(\mathbf{u}_k'\mathbf{Y}, \mathbf{v}_k^{*\prime}\mathbf{X})\cos(\angle(\Sigma_{XY}\mathbf{u}_k, \mathbf{v}_k^o)).\end{aligned} \tag{34}$$

This means that by using OMCPs instead of MCP, we lose $100(1 - |\cos(\angle(\Sigma_{XY}\mathbf{u}_k, \mathbf{v}_k^o))|)\%$ of the optimal covariance. This proportion can be calculated by noting that

$$\cos(\angle(\Sigma_{XY}\mathbf{u}_k, \mathbf{v}_k^o)) = \frac{\mathbf{u}_k'\Sigma_{YX}\mathbf{v}_k^o}{||\mathbf{u}_k'\Sigma_{YX}||}. \tag{35}$$

4. STATISTICAL DOWNSCALING OF TEMPERATURES OVER THE ANTARCTIC

Statistical downscaling is one example where a spatio-temporal process is to be predicted from another. Downscaling refers to inference regarding the behavior of processes at certain spatial scales based on information relevant to processes at comparatively large scales. For example, global numerical weather forecast models or climate models produce model output relevant to physical processes at large scales. In *dynamical* downscaling, these outputs are used to formulate boundary conditions for regional climate models. *Statistical* downscaling involves the use of output from models and observations to build statistical models to predict features at regional or local scales based on large-scale model output. Though we believe that approaches that combine both physical and statistical reasoning are preferred, we focus on statistical downscaling in this article. Examples of statistical downscaling based on dimension-reduction methods including EOFs, MCA, and CCA in the climate literature, for example, von Storch, Zorita, and Cubasch (1993), Gylistras et al. (1994), Cui, von Storch, and Zorita (1995), and Widmann, Bretherton, and Salathè (2003). Overview and comparisons of methods can be found in Wilby et al. (1998) and Zorita and von Storch (1999).

Downscaling climate model outputs is important in the assessment of future impacts of climate change at regional and local scales. There, global climate model outputs, obtained under various scenarios for future controls on the climate system, are used to drive regional or local models. Recent statistical treatments of this problem include Kang and Cressie (2012), Yang and He (2012), and Bürger et al. (2012). For further discussion and references on downscaling, see Mearns et al. (2009).

Here, we apply our dimension-reduced modeling approach to downscale global surface temperature model output to modeled temperatures from a regional model. We do not model true temperatures nor account for model error but focus on the problem of using output from one model to predict results of another model.

4.1 Polar MM5 and ERA-40 Data

Our goal is to downscale ERA-40 reanalysis data to Polar MM5 regional temperatures. The Polar MM5 regional climate model is a fifth-generation Mesoscale Model (MM5) modified for use in polar regions by scientists at the Byrd Polar Research Center at The Ohio State University (Monaghan et al. 2006; Monaghan, Bromwich, and Wang 2006; Monaghan and Bromwich 2008). MM5 was developed at Pennsylvania State University and the National Center for Atmospheric Research. The Polar MM5 simulations were driven by the ERA-40 data and performed on a 121 × 121 polar stereographic grid covering the Antarctic and centered over the South Pole, see Figure 1. The model resolution is 60 km in each horizontal direction. The data we used are seasonally averaged 2 m surface temperatures, available at http://polarmet.osu.edu/PolarMet/ant_hindcast.html.

The global ERA-40 dataset was obtained from the European Centre for Medium-Range Weather Forecasting (ECMWF). These data are the results of the *ERA-40 ReAnalysis Project*. The reanalysis was based on both field observations and satellite data (Uppala et al. 2006). Detailed descriptions of the project are available in the ERA-40 Archive Plan documents from ECMWF (ECMWF 2002). The data used in this article are monthly 2 m temperatures (monthly means of daily means) on a regular 2.5° latitude–longitude grid

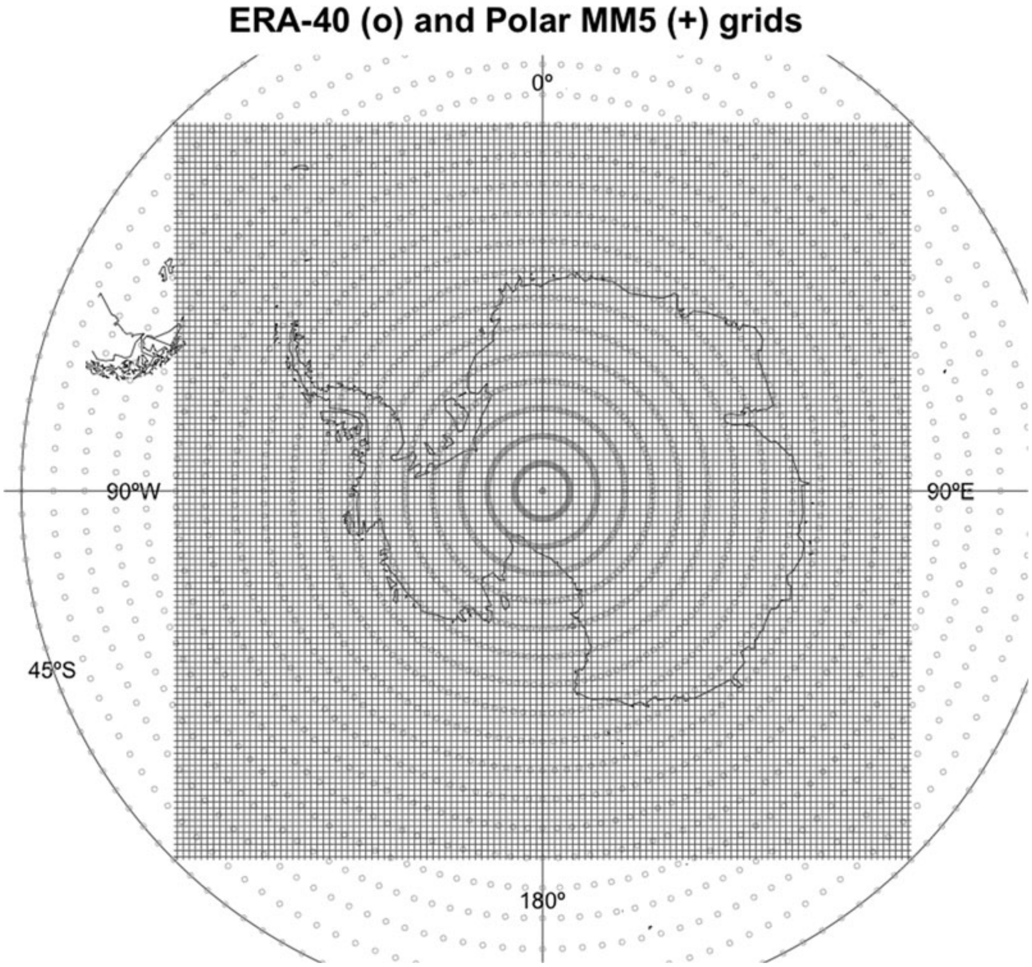

Figure 1. The ERA-40 grid (2736 points) is a 2.5° × 2.5° latitude–longitude grid (○). A stereographic projection of the Polar MM5 grid (14, 641 points) (+).

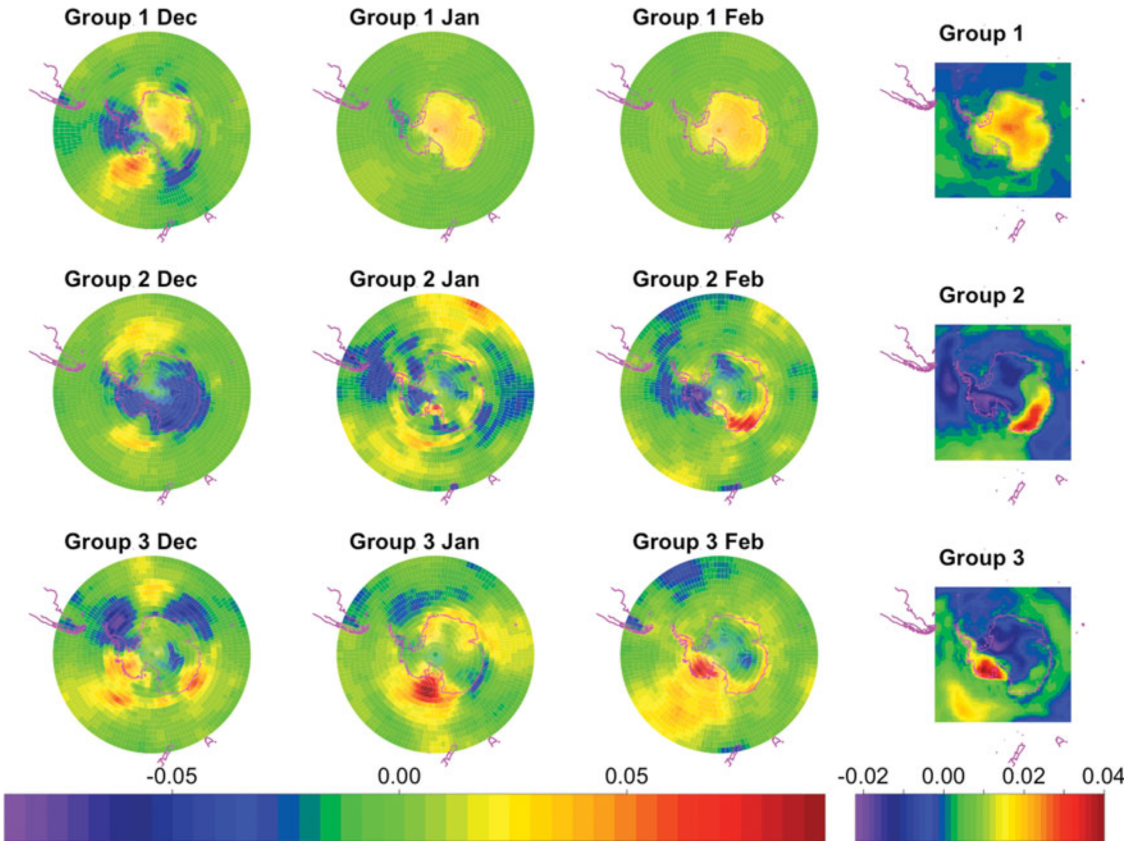

Figure 2. First three groups of OMCPs (circles) and EOF patterns (squares) for summer season.

(see Figure 1), obtained from the ECMWF Data Server (*http://data-portal.ecmwf.int/data/d/era40_moda/*). We only used ERA-40 data south of and including 45°S, which covers the Polar MM5 grid.

4.2 OMCPs and EOF Patterns

We used data for March 1979 through May 2002: a total of 93 seasons yielding more than two million data points. In the development of the dimension reductions, we left out the last year of data so that we can compare them to our predictions for that period. We constructed EOF basis vectors from the centered Polar MM5 data separately for each season.

Using centered ERA-40 data, we obtained OMCPs separately for each month, based on the Polar MM5 EOFs from the corresponding season; see the first three columns in Figure 2. The rationale for developing monthly rather than seasonal OMCP's was two-fold. First, the relationships between monthly temperatures appear stronger than between their seasonal counterparts. Second, this enables prediction of Polar MM5 seasonal results when only one or two of the relevant months of ERA-40 data are available.

The first three groups of EOFs and OMCPs for the Southern Hemisphere summer season are shown in Figure 2. The first group (see the first row of Figure 2) has a strong land versus ocean component. Other patterns often emphasize smaller regions, for example, along the coast of Antarctica or parts of the Pacific Ocean. The proportions of total variance of the Polar MM5 temperature fields explained by increasing number of EOFs for each season are shown in Figure 3. For each season, the first four EOFs capture about 70%–80% of the total variances. With this in mind, we used the first four EOFs and OMCPs in Section 4.3.

4.3 Bayesian Hierarchical Model

Here, we present an overview of the model (details are in the supplementary material).

4.3.1 Data Model. We assumed that the observations are conditionally independent given the amplitude vectors and that

$$\mathbf{Y}_{l,t}|\mathbf{a}_{l,t}, R_l \sim N\left(U_l\mathbf{a}_{l,t}, R_l\right) \text{ and } \mathbf{X}_{m,t}|\mathbf{b}_{m,t}, S_m \sim N\left(V_m\mathbf{b}_{m,t}, S_m\right), \quad (36)$$

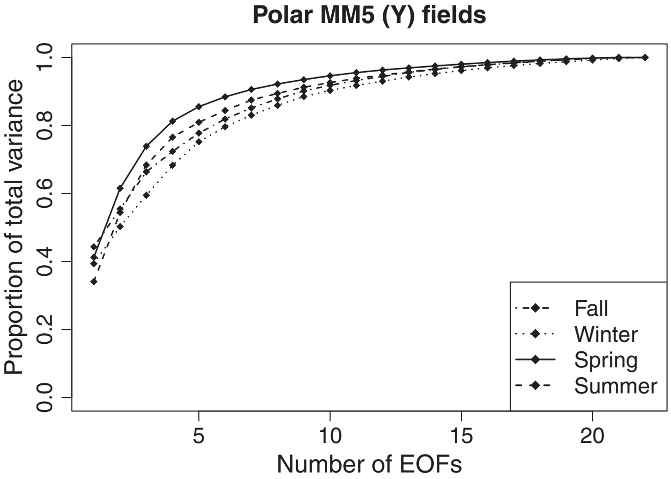

Figure 3. Proportion of total sample variance explained by the EOFs for each season.

where t indicates the year; $\mathbf{Y}_{l,t}$ is the centered Polar MM5 data for season l, $l \in \{1, 2, 3, 4\}$ representing summer, autumn, winter, spring, respectively; $\mathbf{X}_{m,t}$ is the centered ERA-40 data for month m, $m = 1, \dots, 12$, where 1 represents January, etc.; and U_l and V_m contain the first four EOFs and OMCPs for season l and month m. Our treatment of the covariance matrices R_l and S_m is discussed below.

4.3.2 Process Model. We assumed that the Polar MM5 amplitude vectors, $\mathbf{a}_{l,t}$, are conditionally independent given the ERA-40 amplitude vectors, $\mathbf{b}_{m_{l1},t}$, $\mathbf{b}_{m_{l2},t}$, and $\mathbf{b}_{m_{l3},t}$, where the index m_{lk} denotes the kth month within season l and,

$$\mathbf{a}_{l,t}|\mathbf{b}_{m_{l1},t}, \mathbf{b}_{m_{l2},t}, \mathbf{b}_{m_{l3},t}, H_l, C_l \sim N\left(H_l \begin{bmatrix} \mathbf{b}_{m_{l1},t} \\ \mathbf{b}_{m_{l2},t} \\ \mathbf{b}_{m_{l3},t} \end{bmatrix},\ C_l\right), \tag{37}$$

where the C_l are unknown 4×4-dimensional covariance matrices and

$$H_l = \begin{pmatrix} H_{m_{l1}} & H_{m_{l2}} & H_{m_{l3}} \end{pmatrix}, \tag{38}$$

where the H_m are unknown, but based on preliminary data analysis (see the supplementary material), we assumed that the H_m are diagonal.

The $\mathbf{b}_{m,t}$ amplitude vectors were modeled as a first-order Markov process,

$$\mathbf{b}_{m,t}|\mathbf{b}_{m-1,t}, B_m, D_m \sim N\big(B_m\mathbf{b}_{m-1,t},\ D_m\big), \tag{39}$$

where D_m is an unknown 4×4-dimensional covariance matrix and the matrices B_m are unknown. Finally, we used a normal prior on the first ERA-40 amplitude vector, $\mathbf{b}_{2,1979} \sim N\big(\boldsymbol{\mu}_b, \Sigma_b\big)$. Using ERA-40 data for this month, we set $\boldsymbol{\mu}_b = V_2 X_{2,1979} = \hat{\mathbf{b}}_{2,1979} = (-26.784, 12.094, -5.186, 5.562)'$. Estimated standard error of each $\hat{b}_{2,1979,k}$ is 0.72, but to ensure a more vague prior we inflated this value and set $\Sigma_b = 10^2 I_{K_X}$.

4.3.3 Parameter Model. *I. Data model covariances*. Since our primary motivation is modeling in very high dimensions, our treatment of the data model covariance matrices R_l and S_m is critical. Our approach must balance modeling the structure of the data with the computational feasibility. To do so, we applied an approach used in Berliner, Wikle, and Cressie (2000). Let $\tilde{U}_l$ be a matrix that contains columns $K_Y + 1, \dots, L_Y$ of $\mathcal{U}_l$ and let $\tilde{D}_l$ be a diagonal matrix containing the corresponding eigenvalues, $d_{l,j}$ (i.e., the variances associated with these EOFs). As in Berliner, Wikle, and Cressie (2000), we set

$$R_l = r_l\left(c_l I_{N_Y} + \tilde{U}_l \tilde{D}_l \tilde{U}_l'\right) \equiv r_l \tilde{R}_w\ . \tag{40}$$

Here, c_l is a constant set to be the total sample variance left after accounting for the first L_Y EOFs. That is,

$$c_l = \sum_{k=L_Y+1}^{N_Y} d_{l,k} \quad \text{for } l \in \{1, 2, 3, 4\}\ , \tag{41}$$

where $d_{l,k}$ are the eigenvalues of $S_{Y_l} = \frac{1}{T_l - 1} Y_l Y_l'$. Hence, the data model incorporates additional spatial structure beyond that represented in the leading basis vectors used to specify the means. Same approach was used to treat S_m.

These definitions yield significant computational advantages. In particular, the computations never require storage of R_l nor use of matrix multiplication routines. To explain these claims, it is easy to verify that the inverse of R_l defined in (40) is

$$R_l^{-1} = r_l^{-1}\left(c_l^{-1} I_{N_Y} + \tilde{U}_l \tilde{\Lambda}_l \tilde{U}_l'\right), \tag{42}$$

where $\tilde{\Lambda}_l$ is a diagonal matrix with values $\tilde{\lambda}_{l,k} = -d_{l,k}/(c_l(c_l + d_{l,k}))$. This representation leads to simplifications in the Bayesian updating calculations. Those calculations make frequent use of the quantities

$$U_l' R_l^{-1} U_l = (r_l c_l)^{-1} I_{N_Y} \tag{43}$$

and

$$U_l' R_l^{-1} \mathbf{Y}_{l,t} = (r_l c_l)^{-1} U_l' \mathbf{Y}_{l,t}\ . \tag{44}$$

The impact of this formula is that the high-dimensional calculations like the quadratic form in (43) and the first term in (44) can be performed once before the Markov chain Monte Carlo (MCMC) iterations. Further, the memory allocated to the collections U_l and $\tilde{U}_l$ can be cleared before the MCMC begins.

II. Process model parameters. Having little prior information regarding these parameters, we relied on exploratory analysis to guide our selections. To partially mitigate the familiar problems in using the data to form the priors, we selected prior densities that are diffuse compared to the suggestions implied by the data. We assigned normal prior distributions on the elements of H_m and B_m and inverse-Wishart priors on C_l and D_m. Details of hyperparameter selection are in the supplementary material.

4.4 Assessing Conditional Independence Assumptions

Our main assumption of conditional independence is that *given* the EOF amplitudes the Polar MM5 temperatures are independent of the ERA-40 temperatures, that is,

$$[\mathbf{Y}|\mathbf{a}, \mathbf{b}, \boldsymbol{\theta}, \mathbf{X}] = [\mathbf{Y}|\mathbf{a}, \boldsymbol{\theta}], \tag{45}$$

recall the discussion following Equation (12). If we believe that the EOFs capture most of the structure in the Polar MM5 temperatures, this assumption makes intuitive sense. One way to heuristically assess this assumption is to examine the sample correlations between the residuals $\mathbf{R} = \mathbf{Y} - U\mathbf{a}$ and $\mathbf{X}$, or between $\mathbf{R}$ and $\mathbf{X} - V\mathbf{b}$. We have 22 replications of $\mathbf{R}$ and $\mathbf{X}$ for each season–month combination that can be used to calculate the sample correlations between any elements in $\mathbf{R}$ and $\mathbf{X}$. Notice that for each season–month combination there are more than 150 million such pairs so we randomly sampled 10,000 pairs of R_i and X_j; histograms of the sample correlations are shown in Figure 4. To get an idea of what the histograms would look like if there is no correlation we sampled 10,000 pairs of 22-dimensional uncorrelated normal random variables and the solid lines in Figure 4 show a kernel estimate of the resulting "null density." In all cases, the histograms are very close to the null density (except for March), leading us to conclude that the conditional independence assumptions is justified in our example. The same analysis for $\mathbf{R}$ and $\mathbf{X} - V\mathbf{b}$ gave similar results.

4.5 Results

We focus here on the downscaling of ERA-40 data onto the Polar MM5 grid for the time period 2001–2002; recall that the Polar MM5 data for this period were not used in the analysis. We obtained samples from the posterior distribution via Gibbs

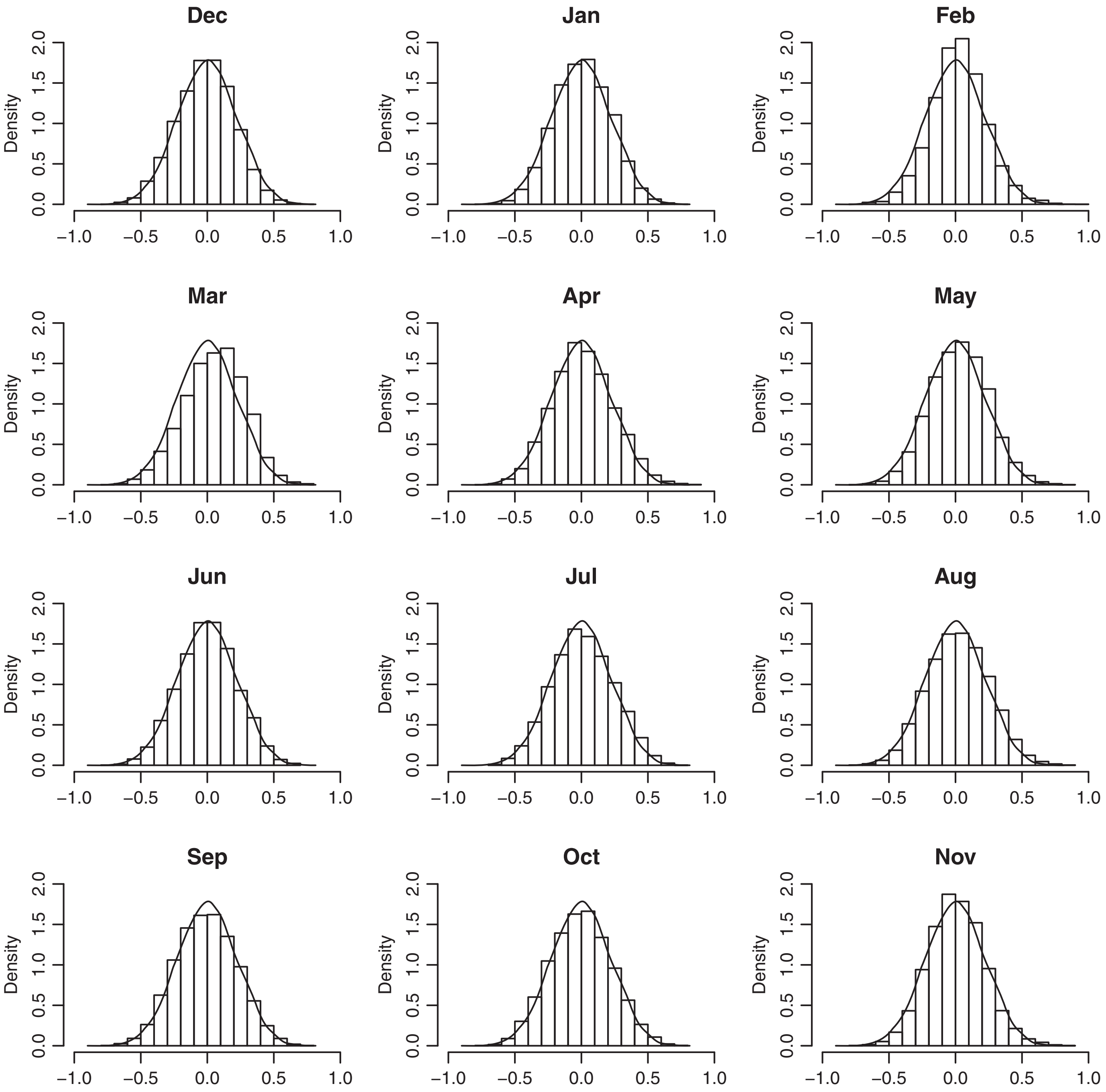

Figure 4. Histograms of sample correlations between 10,000 randomly chosen pairs of residuals $\mathbf{R} = \mathbf{Y} - U\mathbf{a}$ and $\mathbf{X}$. The solid curves show the null density, that is, estimated density of sample correlations between 10,000 uncorrelated pairs of normal random variables.

sampling. The full conditional distributions used by the Gibbs sampler are listed in the supplementary material. We obtained 30,000 MCMC samples and discarded the first 1000 as burn-in. To give a rough idea of computing time, the 30,000 iterations of the Gibbs sampler coded in R took about 4 and one-half hours using an iMac computer with a 2.7 GHz processor and 4GB RAM. We examined trace plots of many of the 1908 parameters (not shown here), which did not indicate any convergence problems.

Posterior predictive inferences for the $\mathbf{a}_{l,t}$ amplitudes for the period 2001–2002 are shown in Figure 5. The 90% prediction intervals are rather wide and most of them cover zero, indicating high predictive uncertainty. For comparison, we also plot the $\hat{a}_{l,t,k}$ amplitudes, estimated from the actual Polar MM5 data from this period. We note that most of them fall within the 90% prediction intervals.

For predictive analysis of the Polar MM5 temperatures, we would ideally obtain samples from the posterior predictive distribution

$$\mathbf{Y}_{l,t} \sim N\left(U_l \mathbf{a}_{l,t}^{(m)}, r_l^{(m)} \tilde{R}_l\right), \tag{46}$$

for each m, where $\mathbf{a}_{l,t}^{(m)}$ and $r_l^{(m)}$ are MCMC samples from the posterior. But sampling this distribution is computationally very expensive due to the high dimensions of $\tilde{R}_l$. However, we can

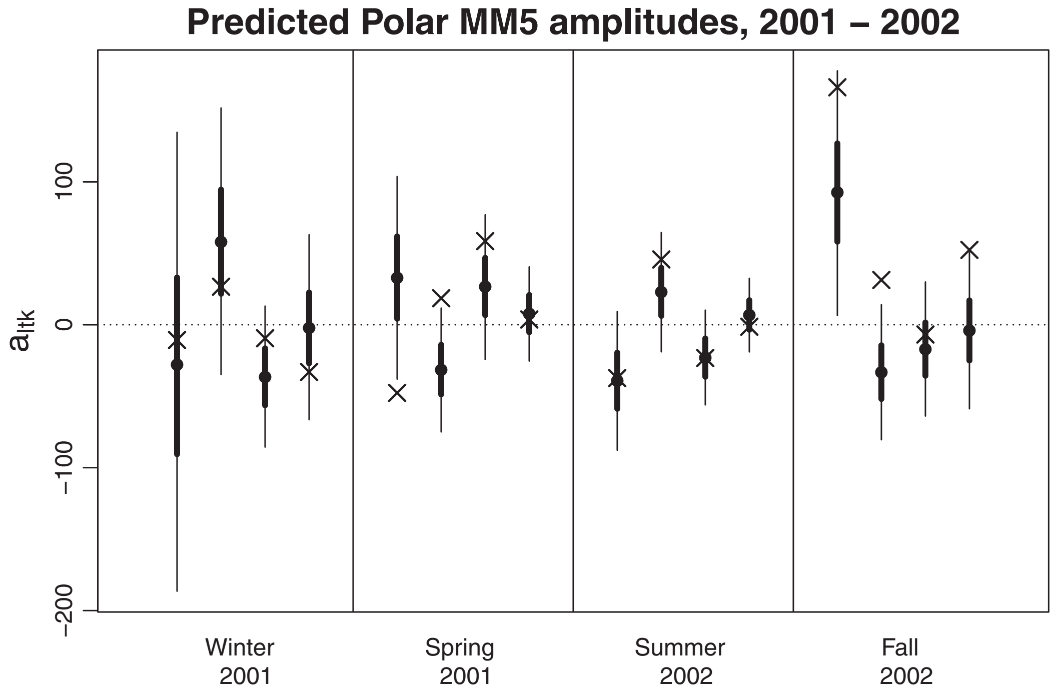

Figure 5. Posterior predictive inferences for the $a_{l,t,k}$ amplitudes for 2001–2002. Circles indicate estimated posterior means. The thick and thin lines show the middle 50% and 90% posterior prediction intervals, respectively. The x-marks show the same amplitudes estimated from the actual Polar MM5 data of this time period.

easily obtain estimates of the predictive mean vector by averaging the $\mathbf{a}_{l,t}^{(m)}$ and multiplying by the corresponding U_l; see Figure 6. Other posterior predictive summaries such as variances and quantiles can be obtained point-wise, using the easily obtained diagonal of $r_l^{(i)}\tilde{R}_l$. Some of these point-wise summaries are shown in Figure 6 along with the actual centered Polar MM5 data from winter 2001 to fall 2002. The downscaled predictive mean temperatures (third column of Figure 6) are expected to be smoother than the observed data but these means still indicate some of the general features in the data. For example, the comparatively cold area in west Antarctica apparent in the summer 2002 Polar MM5 data is also present in the downscaled mean field. The fall 2002 data show warm temperatures over essentially all Antarctica, but the downscaled mean field only shows high temperatures in the west and slightly warmer temperatures in the far east. The standard deviation maps (last column of Figure 6) indicate smaller uncertainties over the ocean than Antarctica. We remark that the ocean temperatures in our dataset do not vary much over the time period in question, making them relatively easy to predict. The 5% and 95% quantiles (second and forth columns of Figure 6) give the impression that the point-wise 90% prediction intervals contain most of the data. To further assess the quality of the predictive model, we present in Table 1 the percentages of data points that fall within the estimated 50%, 90%, and 95% prediction intervals. The observed inclusion percentages are close to the corresponding levels in most cases, indicating that the uncertainty estimation of the predictive model is realistic.

As a check of model adequacy, we obtained point-wise credible intervals for each $Y_{l,t,i}$ and compared with the observations, that is, same approach as in Table 1 but for observations used to fit the model. Overall proportions of observations that fall within the 50% and 90% point-wise credible intervals for $Y_{l,t,i}$ are 0.568 and 0.881, respectively. Figure 7 shows these same proportions for each season separately. The lengths of the point-wise prediction intervals are all less than 0.14 K while the prior standard deviations of $Y_{l,t,i}$ implied by the hierarchical model are in the order of 30 K (see the supplementary material) showing that reduction of uncertainty from prior to posterior is substantial. Overall the model seems to fit the data well.

In Section 3.2, we gave theoretical arguments why we prefer MCP (or OMCP) vectors for the **X** process over EOF patterns. First, that MCPs maximize covariance between amplitudes and second, the proportion of total variation in the target process **Y** is higher when using (the first) MCP for the **X** process versus (the first) EOF. These arguments hold on average, but to compare the performance of OMCPs and EOFs in this particular application, we repeated the analysis where the first four OMCPs were replaced by the first four EOFs. To compare the predictive performance, we calculated the mean square prediction error (MSPE) for the last year, using the posterior predicted means as estimators, see Table 2. The results for these data are mixed, OMCPs do better for Winter and Summer while EOFs do

Table 1. Percentages of locations where the Polar MM5 temperature data fall within the 50%, 90%, and 95% prediction intervals

	Winter 2001	Spring 2001	Summer 2002	Fall 2002
50% Prediction interval	70.72%	38.19%	52.52%	50.67%
90% Prediction interval	97.40%	77.58%	89.39%	80.60%
95% Prediction interval	98.60%	86.45%	94.58%	85.31%

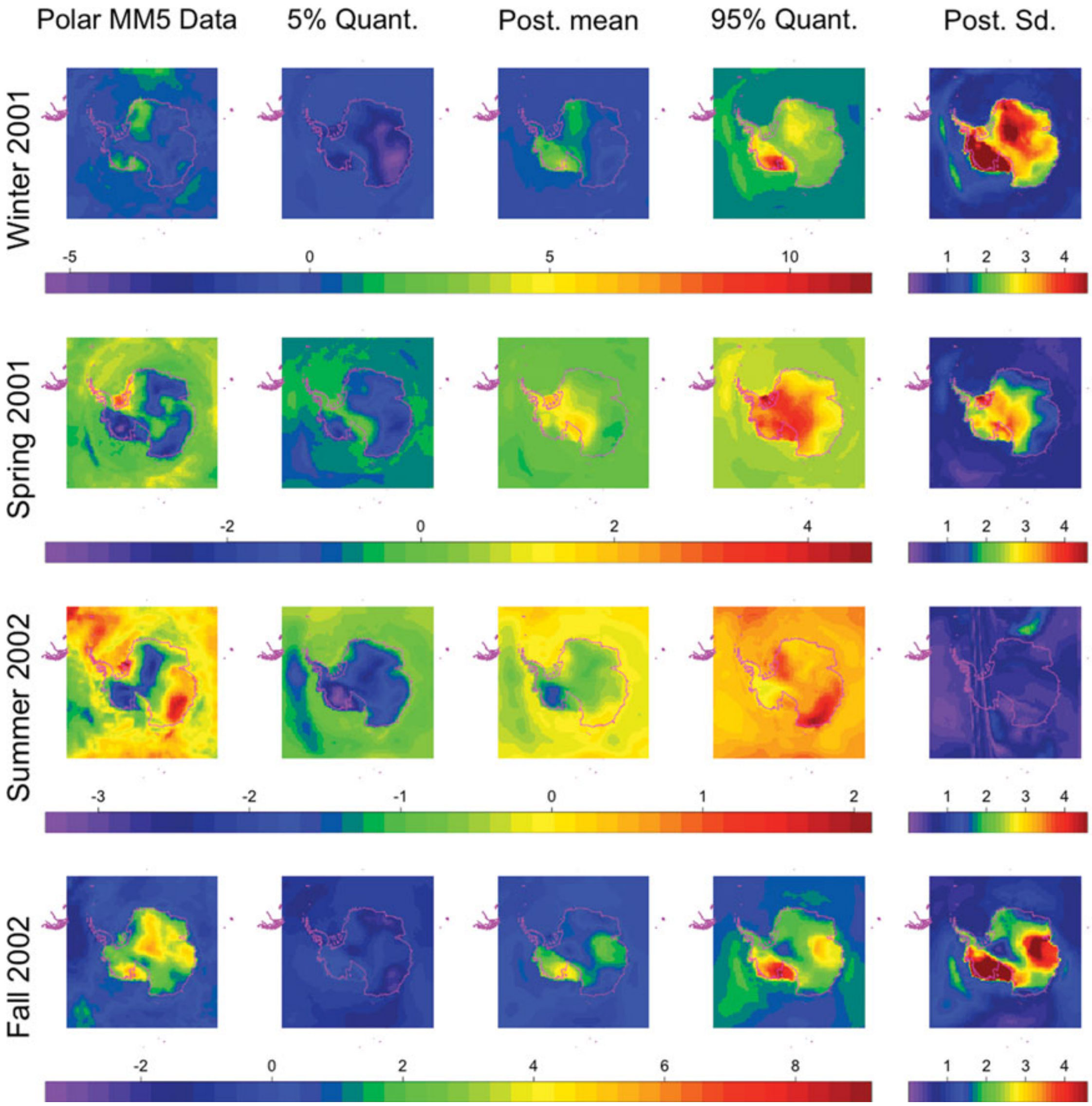

Figure 6. First column: the actual Polar MM5 temperature data winter 2001 to fall 2002. The remaining columns show summaries of the posterior predictive distribution of Polar MM5 temperature fields: 5% quantile, posterior mean, 95% quantile and the standard deviation.

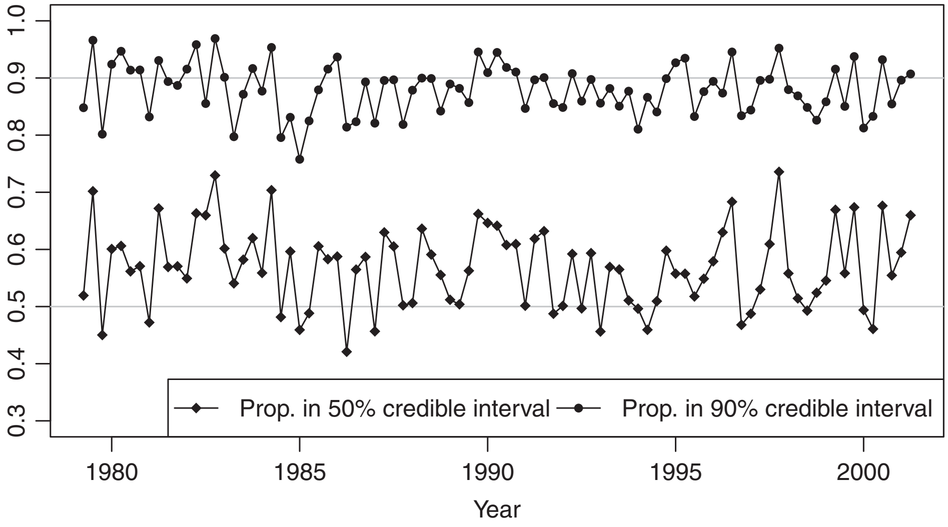

Figure 7. Proportions of Polar MM5 temperatures that fall within estimated 50% and 90% credible intervals for $Y_{l,t,i}$.

Table 2. Mean square prediction error of the Polar MME temperatures using OMCP vectors versus EOF vectors for the ERA-40 temperatures

	Winter 2001	Spring 2001	Summer 2002	Fall 2002
OMCP	**0.420**	0.947	**0.266**	1.533
EOF	0.826	**0.912**	0.275	**1.331**

better for Spring and Fall (shown in bold in Table 2). Table 3 shows the percentages of data points that fall within the estimated prediction intervals. The results are comparable with Table 1, the EOFs results are slightly less over-confident for Spring and Fall but more conservative for Winter and Summer.

5. DISCUSSION

We have outlined a general Bayesian hierarchical approach to dimension-reduced modeling of two spatio-temporal processes. We discussed several aspects of the general framework in Section 2. Furthermore, we have proposed methods of obtaining data-dependent basis vectors, namely, MCPs and OMCPs (Section 3). We have applied these methods to statistical downscaling of surface temperatures in the Antarctic (Section 4). We close with a brief discussion of a few points.

Our goal has been to predict one spatio-temporal process (**Y**) from another (**X**), using dimension-reduced representations of both processes. The motivation for using the dimension-reduction approach is the need to be able to handle large datasets where a full estimation of the statistical properties of the processes is not feasible due to computational constraints. Of course, a dimension-reduced approach may lead to suboptimal predictors. We partially addressed this issue in Section 2.1. Assuming a linear model between two spatial processes, we investigated the difference in MSE between the optimal predictor and predictors based on our dimension-reduced approach. We demonstrated that the increase in MSE depends on how well the dimension-reduced model captures the full linear model (see Equation (26)). Furthermore, we showed that the MSE of a dimension-reduced predictor depends on how well the dimension-reduced representation captures the target process and the MSE of the predictor of the amplitudes (see Equation (24)).

We developed and suggest the use of MCPs or OMCPs for the **X** process because they are direct approaches to the prediction problem. In Section 4, we used EOFs for the **Y** process because of their property of maximizing the proportion of total variance explained, but any other set of basis vectors could be considered. For example, if the target process is a combination of known physical patterns that we wish to predict, we can do so with MCPs or OMCPs. If the only goal is prediction, MCPs are likely to perform better than OMCPs. However, OMCPs can be chosen over MCPs due to their computational advantages discussed in Section 4.3.3.

Table 3. Percentages of locations where the Polar MM5 temperature data fall within the 50%, 90%, and 95% prediction intervals from the analysis where EOF basis vectors are used for ERA-40 temperatures

	Winter 2001	Spring 2001	Summer 2002	Fall 2002
50% Prediction interval	62.39%	42.33%	56.20%	54.98%
90% Prediction interval	97.68%	84.88%	91.90%	88.07%
95% Prediction interval	98.45%	92.58%	95.75%	92.10%

The statistical downscaling example in Section 4 demonstrated the effectiveness of the hierarchical dimension-reduced modeling approach for two spatio-temporal processes. The Polar MM5 and ERA-40 are both very large fields so direct modeling approaches would be computationally very difficult. The dimension-reduced hierarchical model produced satisfactory results when downscaling ERA-40 data onto the Polar MM5 grid. We also fitted a model where both the Polar MM5 and ERA-40 data of the last year were left out. In that case, we first predicted the **b** amplitude vectors and then the **a** amplitude vectors and lastly the prediction of the Polar MM5 field. The results were not great, the model failed to adequately capture the temporal evolution of the ERA-40 fields. In essence, the OMCPs of ERA-40 data were very effective for predicting Polar MM5 data (i.e., downscaling), but they were not effective in capturing the temporal evolution of the ERA-40 data. This is perhaps not surprising since the focus of the OMCPs is on the spatial variation of the data but not the temporal variation. This same concern is true for EOFs, see, for example, sec. 7.3.1 in Cressie and Wikle (2011), but as a tool for predicting one process from another the OMCPs and MCPs are very promising.

SUPPLEMENTARY MATERIALS

The online supplementary material provides more details regarding the application in Section 4. There we present more exploratory data analysis, a detailed description of the Bayesian model and prior selection, address the issue of model adequacy and outline the Gibbs sampler we used.

[Received December 2012. Revised March 2014.]

REFERENCES

Berliner, L. M. (1996), “Hierarchical Bayesian Time Series Models,” in *Maximum Entropy and Bayesian Methods*, eds. K. Hanson and R. Silver, Dordrecht, the Netherlands: Kluwer Academic Publishers, pp. 15–22. [1649]

——— (2003), “Physical-Statistical Modeling in Geophysics,” *Journal of Geophysical Research*, 108, STS31–10. [1648]

Berliner, L. M., Wikle, C. K., and Cressie, N. (2000), “Long-Lead Prediction of Pacific SSTs via Bayesian Dynamic Modeling,” *Journal of Climate*, 13, 3953–3968. [1648,1650,1654]

Bürger, G., Murdock, T. Q., Werner, A. T., Sobie, S. R., and Cannon, A. J. (2012), “Downscaling Extremes—An Intercomparison of Multiple Statistical Methods for Present Climate,” *Journal of Climate*, 25, 4366–4388. [1652]

Calder, C. A. (2007), “Dynamic Factor Process Convolution Models for Multivariate Space–Time Data With Application to Air Quality Assessment,” *Environmental and Ecological Statistics*, 14, 229–247. [1648]

Cressie, N., Shi, T., and Kang, E. L. (2010), “Fixed Rank Filtering for Spatio-Temporal Data,” *Journal of Computational and Graphical Statistics*, 19, 724–745. [1648]

Cressie, N., and Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data*, Hoboken, NJ: Wiley. [1647,1648,1650,1658]

Cui, M., von Storch, H., and Zorita, E. (1995), “Coastal Sea Level and the Large-Scale Climate State: A Downscaling Exercise for the Japanese Islands,” *Tellus*, 47, 132–144. [1652]

Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R., and Palsson, B. Ø. (2007), "Global Reconstruction of the Human Metabolic Network Based on Genomic and Bibliomic Data," *Proceedings of the National Academy of Sciences*, 104, 1777–1782. [1647]

ECMWF, (2002), *The ERA-40 Archive* [On-line version available from the ECMWF ERA-40 Project Plan], Reading, United Kingdom: European Centre for Medium-Range Weather Forecasts. Available at *http://www.ecmwf.int/research/era/Project/Plan/* [1652]

Esteban-Parra, M. J., Rodrigo, F. S., and Castro-Diez, Y. (1998), "Spatial and Temporal Patterns of Precipitation in Spain for the Period 1880–1992," *International Journal of Climatology*, 18, 1557–1574. [1650]

Golub, G. H., and Van Loan, C. F. (1996), *Matrix Computation* (3rd ed.), Baltimore, MD: Johns Hopkins University Press. [1650]

Gylistras, D., von Storch, H., Fischlin, A., and Beniston, M. (1994), "Linking GCM-Simulated Climatic Changes to Ecosystem Models: Case Studies of Statistical Downscaling in the Alps," *Climate Research*, 4, 167–189. [1652]

Higdon, D. (1998), "A Process-Convolution Approach to Modelling Temperatures in the North Atlantic Ocean," *Environmental and Ecological Statistics*, 5, 173–190. [1648]

Johnson, R. A., and Wickern, D. W. (2002), *Applied Multivariate Statistics* (5th ed.), New Jersey: Prentice Hall. [1650]

Kang, E. L., and Cressie, N. (2012), "Bayesian Hierarchical ANOVA of Regional Climate-Change Projections From NARCCAP Phase II," *International Journal of Applied Earth Observation and Geoinformation*, 22, 3–15. [1652]

Kang, E. L., Cressie, N., and Shi, T. (2010), "Using Temporal Variability to Improve Spatial Mapping With Application to Satellite Data," *The Canadian Journal of Statistics*, 38, 271–289. [1648]

Katzfuss, M., and Cressie, N. (2011), "Spatio-Temporal Smoothing and EM Estimation for Massive Remote-Sensing Data Sets," *Journal of Time Series Analysis*, 32, 430–446. [1648]

——— (2012), "Bayesian Hierarchical Spatio-Temporal Smoothing for Very Large Datasets," *Environmetrics*, 23, 94–107. [1648]

Kleiber, W., Sain, S. R., Heaton, M. J., Wiltberger, M., Reese, C. S., and Bingham, D. (2013), "Parameter Tuning for a Multi-Fidelity Dynamical Model of the Magnetosphere," *Annals of Applied Statistics*, 7, 1286–1310. [1647]

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009), "Computational Social Science," *Science*, 323, 721–723. [1647]

Mearns, L. O., Gutowski, W. J., Jones, R., Leung, L.-Y., McGinnis, S., Nunes, A. M. B., and Qian, Y. (2009), "A Regional Climate Change Assessment Program for North America," *Eos, Transactions, American Geophysical Union*, 90, 311–312. [1652]

Monaghan, A. J., and Bromwich, D. H. (2008), "Advances in Describing Recent Antarctic Climate Variability," *Bulletin of the American Meteorological Society*, 89, 1295–1306. [1652]

Monaghan, A. J., Bromwich, D. H., Fogt, R. L., Wang, S.-H., Mayewski, P. A., Dixon, D. A., Ekaykin, A., Frezzotti, M., Goodwin, I., Isaksson, E., Kaspari, S. D., Morgan, V. I., Oerter, H., Van Ommen, T. D., Van der Veen, C. J., and Wen, J. (2006), "Insignificant Change in Antarctic Snowfall Since the International Geophysical Year," *Science*, 313, 827–831. [1652]

Monaghan, A. J., Bromwich, D. H., and Wang, S. H. (2006), "Recent Trends in Antarctic Snow Accumulation From Polar MM5 Simulations," *Philosophical Transactions A Mathematical, Physical, and Engineering Sciences*, 364, 1683–1708. [1652]

Rao, C. R. (1973), *Linear Statistical Inference and its Applications* (2nd ed.), New York: Wiley. [1649]

Ravikumar, P., Vu, V. Q., Yu, B., Naselaris, T., Kay, K., and Gallant, J. (2009), "Nonparametric Sparse Hierarchical Models Describe v1 Fmri Responses to Natural Images," in *Advances in Neural Information Processing Systems 21*, eds. D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, NIPS Foundation, Red Hook, NY: Curran Associates Incorporated, pp. 1337–1344. [1647]

Richardson, S., Bottolo, L., and Rosenthal, J. S. (2010), "Bayesian Models for Sparse Regression Analysis of High Dimensional Data," in *Bayesian Statistics 9*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Oxford: Oxford University Press. [1647]

Royle, J. A., and Berliner, L. M. (1999), "A Hierarchical Approach to Multivariate Spatial Modeling and Prediction," *Journal of Agricultural, Biological, and Environmental Statistics*, 4, 29–56. [1648]

Stroud, J. R., Müller, P., and Sansó, B. (2001), "Dynamic Models for Spatiotemporal Data," *Journal of the Royal Statistical Society*, Series B, 63, 673–689. [1648]

Uppala, S. M., Kållberg, P. W., Simmons, A. J., Andrae, U., Da Costa Bechtold, V., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Van De Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J. (2006), "The Era-40 Re-Analysis," *Quarterly Journal of the Royal Meteorological Society*, 131, 2961–3012. [1652]

Ver Hoef, J. M., and Barry, R. P. (1998), "Constructing and Fitting Models for Cokriging and Multivariable Spatial Prediction," *Journal of Statistical Planning and Inference*, 69, 275–294. [1648]

von Storch, H., Zorita, E., and Cubasch, U. (1993), "Downscaling of Global Climate Change Estimates to Regional Scales: An Application to Iberian Rainfall in Wintertime," *Journal of Climate*, 6, 1161–1171. [1652]

von Storch, H., and Zwiers, F. W. (2001), *Statistical Analysis in Climate Research*, Cambridge: Cambridge University Press (paperback (with corrections) edition). [1650]

Widmann, M., Bretherton, C. S., and Salathè, E. P. (2003), "Statistical Precipitation Downscaling Over the Northwestern United States Using Numerically Simulated Precipitation as a Predictor," *Journal of Climate*, 16, 799–816. [1652]

Wikle, C. K. (2010), "Low-Rank Representations for Spatial Processes," in *Handbook of Spatial Statistics*, eds. A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp, Boca Raton, FL: Chapman & Hall/CRC, pp. 107–118. [1648]

Wikle, C. K., and Cressie, N. (1999), "A Dimension-Reduced Approach to Space-Time Kalman Filtering," *Biometrika*, 86, 815–829. [1648]

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. (2001), "Spatiotemporal Hierarchical Bayesian Modeling: Tropical Ocean Surface Winds," *Journal of the American Statistical Association*, 96, 382–397. [1648]

Wilby, R. L., Wigley, T. M. L., Conway, D., Jones, P. D., Hewitson, B. C., Main, J., and Wilks, D. S. (1998), "Statistical Downscaling of General Circulation Model Output: A Comparison of Methods," *Water Resources Research*, 34, 2995–3008. [1652]

Xu, K., Wikle, C. K., and Fox, N. I. (2005), "A Kernel-Based Spatio-Temporal Dynamical Model for Nowcasting Weather Radar Reflectivities," *Journal of the American Statistical Association*, 100, 1133–1144. [1648]

Yang, Y., and He, X. (2012), "Bayesian Empirical Likelihood for Quantile Regression," *The Annals of Statistics*, 40, 1102–1131. [1652]

Zhang, Y., Wallace, J. M., and Battisti, D. S. (1997), "ENSO-Like Interdecadal Variability: 1900–93," *Journal of Climate*, 10, 1004–1020. [1650]

Zorita, E., and von Storch, H. (1999), "The Analog Method as a Simple Statistical Downscaling Technique: Comparison With More Complicated Methods," *Journal of Climate*, 12, 2472–2489. [1652]