J Geogr Syst (2012) 14:29–47
DOI 10.1007/s10109-011-0154-8

ORIGINAL ARTICLE

Bayesian dynamic modeling for large space-time datasets using Gaussian predictive processes

Andrew O. Finley · Sudipto Banerjee · Alan E. Gelfand

Received: 18 March 2011 / Accepted: 26 July 2011 / Published online: 26 August 2011

Abstract In this paper, we extend the applicability of a previously proposed class of dynamic space-time models by enabling them to accommodate large datasets. We focus on the common setting where space is viewed as continuous but time is taken to be discrete. Scalability is achieved by using a low-rank *predictive process* to reduce the dimensionality of the data and ease the computational burden of estimating the spatio-temporal process of interest. The proposed models are illustrated using weather station data collected over the northeastern United States between 2000 and 2005. Here our interest is to use readily available predictors, association among measurements at a given station, as well as dependence across space and time to improve prediction for incomplete station records and locations where station data does not exist.

Keywords Bayesian inference · Dynamic models · Spatial processes · Predictive process

JEL Classification C · Q

Electronic supplementary material The online version of this article (doi:10.1007/s10109-011-0154-8) contains supplementary material, which is available to authorized users.

A. O. Finley (✉)
Departments of Geography and Forestry, Michigan State University, East Lansing, MI, USA
e-mail: finleya@msu.edu

S. Banerjee
Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, USA
e-mail: baner009@umn.edu

A. E. Gelfand
Department of Statistical Science, Duke University, Durham, NC, USA
e-mail: alan@stat.duke.edu

1 Introduction

The scientific community is moving into an era where data-rich environments provide extraordinary opportunities to understand the complexity of spatially and temporally dynamic systems. Unprecedented access to data is a result of investments to collect information for regulatory, monitoring, and resource management objectives, as well as technological advances in spatially enabled sensor networks, and geospatial information storage, analysis, and distribution systems. Across scientific fields, researchers face the challenge of coupling these data with imperfect models to better understand variability in their system of interest. The inference garnered through these analyses often supports decisions with important economic and ecological implications, and it is therefore critical to accurately assess inferential uncertainty. The obstacle for researchers is increasingly often not access to the right data, but rather to implement appropriate statistical methods and software. However, developing modeling frameworks capable of accounting for various sources of uncertainty is not a trivial task—massive datasets, missing observations, and complex dependence structures only serve to exacerbate these challenges.

There is a considerable literature in spatio-temporal modeling. The approach adopted here applies to the setting where space is viewed as continuous, but time is taken to be discrete. Put another way, we view the data as a time series of spatial process realizations and work in the setting of dynamic models, achieving a class of dynamic models for such data. Building upon previous work in the setting of dynamic models by West and Harrison (1997), several authors, including Tonellato (1997), Stroud et al. (2001) and Gelfand et al. (2005), proposed dynamic frameworks to model residual spatial and temporal dependence. These proposed frameworks are flexible and easily extended to accommodate nonstationary and multivariate outcomes. However, full inference and accurate assessment of uncertainty of the resulting model involve computation of order $O(n^3)$ for each time point, where n is the number of locations. This renders the models computationally infeasible for datasets beyond perhaps a few hundred locations and a few time steps. For example, Gelfand et al. (2005) analyzed a space-time dataset of observations from only 50 spatial locations monitored over 12 months.

In this paper, we detail a modeling approach which allows a class of dynamic space-time models proposed by Gelfand et al. (2005) to be applied to large spatio-temporal datasets without sacrificing model flexibility and with minimal loss of information. We focus on the common setting where space is considered continuous but time is taken to be discrete. Some examples of data that fit this description include: US Environmental Protection Agency's Air Quality System which reports pollutants' mean, minimum, and maximum at 8 and 24 h intervals; climate model outputs of weather variables generated on hourly or daily intervals; and remotely sensed landuse/landcover change recorded at annual or decadal time steps.

Modeling large spatial datasets has received much attention in the statistical literature. Vecchia (1988) proposed approximating the likelihood with a product of appropriate conditional distributions to obtain maximum likelihood estimates. Stein et al. (2004) adapt this to restricted maximum likelihood estimation. Another

possibility is to approximate the likelihood using spectral representations of the spatial process (Fuentes 2007). These likelihood approximations yield a joint distribution, but not a process that facilitates spatial interpolation. Yet another approach considers compactly supported correlation functions (Furrer et al. 2006; Kaufman et al. 2008; Du et al. 2009) that yield sparse correlation matrices which, given sparse solvers, are used for computationally efficient kriging and variance estimation. This approach, however, does sacrifice some information and may limit modeling flexibility as full likelihood-based inference still requires determinant computations that may be problematic. Also recognizing this challenge, Pace and LeSage (2009) detail a sampling-based approach to estimating the determinant of massive spatial covariance matrices.

Rather than approximations, one could build models especially geared toward handling of large spatial datasets. These are representations of the spatial process in a lower-dimensional subspace and are often referred to as low-rank or reduced-rank spatial models (Higdon 2002; Kamman and Wand 2003; Stein 2007, 2008; Cressie and Johannesson 2008; Banerjee et al. 2008; Crainiceanu et al. 2008). The idea here is to consider a smaller set of points, often referred to as "knots," where the number of knots is *fixed* to be much smaller than the number of observed locations, and to express the spatial process in terms of its realizations over the smaller set of knots. It is reasonable to assume there will be insignificant loss of spatial information in the underlying process from using a smaller set of locations that determines a subspace to project the data onto. Subsequently, we consider a special class of low-rank processes called the *predictive process* (Banerjee et al. 2008). This arises from a conditional expectation of the original process (referred to as the *parent process*) given its realization over the knots. As such, the predictive process model achieves certain optimality properties (Banerjee et al. 2008) and, also, is a dimension reduction technique that requires no additional tuning parameters in the modeling.

As in Gelfand et al. (2005), we adopt a Bayesian framework to fit the proposed hierarchical models. Here, inference about model parameters and subsequent prediction at new locations and times is based on samples from posterior distributions. These samples are collected using Markov chain Monte Carlo (MCMC) methods. The statistical literature acknowledges that spatial and temporal associations are captured most effectively using models that build dependencies in different stages (Banerjee et al. 2004). Hierarchical models are especially advantageous with datasets having several lurking sources of uncertainty and dependence, where they enable much richer models with less stringent assumptions than traditional modeling paradigms (see e.g., Carlin and Louis 2000; Gelman et al. 2004).

The remainder of the paper evolves as follows. Section 2 defines the proposed predictive process for spatio-temporal modeling and an efficient MCMC sampling and prediction algorithm. Here too, we offer a criterion to assess model fit. In Sect. 3, we explore the proposed model using an analysis of weather station data collected over the northeastern United States between 2000 and 2005. Finally, Sect. 4 concludes the paper with a brief summary and description of future work.

2 Dynamic spatio-temporal models

Dynamic linear models, or state-space models, have gained tremendous popularity in recent years in fields as disparate as engineering, economics, genetics, and ecology. They offer a versatile framework for fitting several time-varying models (West and Harrison 1997). Gelfand et al. (2005) adapted the dynamic modeling framework to spatio-temporal models with spatially varying coefficients. Alternative adaptations of dynamic linear models to space-time data can be found in Stroud et al. (2001) and Tonellato (1997). Our contribution is to present a low-rank version of such models using the predictive process (Banerjee et al. 2008), allowing us to analyze much larger datasets.

Suppose, $y_t(\boldsymbol{s})$ denotes the observation at location $\boldsymbol{s}$ and time t. We model $y_t(\boldsymbol{s})$ through a *measurement equation* that provides a regression specification with a space-time varying intercept and serially and spatially uncorrelated zero-centered Gaussian disturbances as measurement error $\epsilon_t(\boldsymbol{s})$. Next a *transition equation* introduces a $p \times 1$ coefficient vector, say $\boldsymbol{\beta}_t$, which is a purely temporal component (i.e., time-varying regression parameters), and a spatio-temporal component $u_t(\boldsymbol{s})$. Both these are generated through transition equations, capturing their Markovian dependence in time. While the transition equation of the purely temporal component is akin to usual state-space modeling, the spatio-temporal component is generated using Gaussian spatial processes. The overall model is written as

$$\begin{aligned} y_t(\boldsymbol{s}) &= \boldsymbol{x}_t(\boldsymbol{s})'\boldsymbol{\beta}_t + u_t(\boldsymbol{s}) + \epsilon_t(\boldsymbol{s}), \quad \epsilon_t(\boldsymbol{s}) \overset{\text{ind.}}{\sim} N(0, \tau_t^2) \\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \overset{\text{i.i.d.}}{\sim} N(0, \boldsymbol{\Sigma}_\eta) \\ u_t(\boldsymbol{s}) &= u_{t-1}(\boldsymbol{s}) + w_t(\boldsymbol{s}), \quad w_t(\boldsymbol{s}) \overset{\text{ind.}}{\sim} GP(\boldsymbol{0}, C_t(\cdot, \boldsymbol{\theta}_t)), \quad t = 1, 2, \ldots, N_t, \end{aligned} \tag{1}$$

where the abbreviations ind. and i.i.d. are *independent* and *independent and identically distributed*, respectively. Here $\boldsymbol{x}_t(\boldsymbol{s})$ is a $p \times 1$ vector of predictors and $\boldsymbol{\beta}_t$ is a $p \times 1$ vector of coefficients. In addition to an intercept, $\boldsymbol{x}_t(\boldsymbol{s})$ can include location-specific variables useful for explaining the variability in $y_t(\boldsymbol{s})$. The $GP(\boldsymbol{0}, C_t(\cdot, \boldsymbol{\theta}_t))$ denotes a spatial Gaussian process (a Gaussian process defined over an Euclidean spatial domain; see, e.g., Cressie 1993) with covariance function $C_t(\cdot; \boldsymbol{\theta}_t)$. We customarily specify $C_t(\boldsymbol{s}_1, \boldsymbol{s}_2; \boldsymbol{\theta}_t) = \sigma_t^2 \rho(\boldsymbol{s}_1, \boldsymbol{s}_2; \phi_t)$, where $\boldsymbol{\theta}_t = \{\sigma_t^2, \phi_t\}$ and $\rho(\cdot; \phi)$ are a *correlation function* with ϕ controlling the correlation decay and σ_t^2 represents the spatial variance component. An exponential function is often used to define the spatial correlation structure, e.g., $C_t(\boldsymbol{s}_1, \boldsymbol{s}_2; \boldsymbol{\theta}_t) = \sigma_t^2 \exp(-\phi_t \|\boldsymbol{s}_1 - \boldsymbol{s}_2\|)$, where $\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|$ is the Euclidean distance between the sites $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. However, any *valid* spatial correlation function could be used, see, e.g., Cressie 1993, Chilés and Delfiner 1999, and Banerjee et al. 2004. We further assume $\boldsymbol{\beta}_0 \sim N(\boldsymbol{m}_0, \boldsymbol{\Sigma}_0)$ and $u_0(\boldsymbol{s}) \equiv 0$, which completes the prior specifications leading to a well-identified Bayesian hierarchical model and also yield reasonable dependence structures. In practice, estimation of model parameters is usually very robust to these hyper-prior specifications. Also note that (1) reduces to a simple spatial regression model for $t = 1$.

We consider settings where the inferential interest lies in spatial prediction or interpolation over a region for a set of discrete time points. We also assume that the

same locations are monitored for each time point resulting in a space-time matrix whose rows index the locations and columns index the time points, i.e., the (i, j)-th element is $y_j(\boldsymbol{s}_i)$. Our algorithm will accommodate the situation where some cells of the space-time data matrix may have missing observations, as is common in monitoring pollutants or weather variables. In particular, we focus upon situations where the number of monitored locations is large enough to preclude estimation of (1) but the number of time points is manageable (i.e., the space-time data matrix has a large number of rows).

Conducting full Bayesian inference for (1) is computationally onerous. One usually seeks to sample from the posterior distribution of the parameters (including the spatial-temporal random effects) using Gibbs sampling. Implementing this algorithm requires matrix decompositions of cubic order in the number of locations for each time point—and this is repeated in each iteration of the Gibbs sampler.

To circumvent this computational bottleneck, we replace the spatial process in (1) with a low-rank process called the predictive process (Banerjee et al. 2008). More precisely, we consider a smaller set of sites in the domain of interest called "knots," say $\mathcal{S}^* = \{\boldsymbol{s}_1^*, \boldsymbol{s}_2^*, \ldots, \boldsymbol{s}_{n^*}^*\}$ with $n^* \ll n$, and write $\tilde{w}_t(\boldsymbol{s}) = E[w_t(\boldsymbol{s})|\boldsymbol{w}_t^*]$, where $\boldsymbol{w}_t^* = (w_t(\boldsymbol{s}_1^*), w_t(\boldsymbol{s}_2^*), \ldots, w_t(\boldsymbol{s}_{n^*}^*))'$. We obtain a predictive process counterpart of (1) by replacing $u_t(\boldsymbol{s})$ in (1) with $\tilde{u}_t(\boldsymbol{s}) = \sum_{k=1}^{t}[\tilde{w}_k(\boldsymbol{s}) + \tilde{\epsilon}_k(\boldsymbol{s})]$, where the "adjustment" $\tilde{\epsilon}_k(\boldsymbol{s})$ compensates for the oversmoothing by the conditional expectation component and the consequent underestimation of spatial variability (Finley et al. 2009). To be precise, we take $\tilde{\epsilon}_k(\boldsymbol{s}) \overset{\text{ind.}}{\sim} N(0, \delta_k^2(\boldsymbol{s}))$ with $\delta_k^2(\boldsymbol{s}) = C_k(\boldsymbol{s}, \boldsymbol{s}; \boldsymbol{\theta}_k) - \boldsymbol{c}_k(\boldsymbol{s}; \boldsymbol{\theta}_k)' \boldsymbol{C}_k^*(\boldsymbol{\theta}_k)^{-1} \boldsymbol{c}_k(\boldsymbol{s}; \boldsymbol{\theta}_k)$, where $\boldsymbol{c}_k(\boldsymbol{s}; \boldsymbol{\theta}_k)'$ is a $1 \times n^*$ vector whose j-th element is $C_k(\boldsymbol{s}, \boldsymbol{s}_j^*; \boldsymbol{\theta}_k)$ and $\boldsymbol{C}_k^*(\boldsymbol{\theta}_k)$ is the $n^* \times n^*$ matrix whose (i, j)-th element is $C_k(\boldsymbol{s}_i^*, \boldsymbol{s}_j^*; \boldsymbol{\theta}_k)$. Moving to the predictive process model immediately reduces computational complexity from $O(n^3)$ to $O(n^{*3})$ and hence reduces run-time.

A key issue in this and similar low-rank models is the choice of knots. Given a computationally feasible n^* one could fix the knot location using grid over the extent of the domain, space-covering design (e.g., Royle and Nychka 1998), or more sophisticated approach aimed at minimizing a predictive variance criterion (see, e.g., Finley et al. 2009; Guhaniyogi et al. 2011). In practice, if the observed locations are evenly distributed across the domain, we have found relatively small difference in inference based on knot locations chosen using a grid, space-covering design, or other criterion. Rather, it is the number of knots locations that has the greater impact on parameter estimates and subsequent prediction. Therefore, we often investigate sensitivity of inference to different knot intensities, within a computationally feasible range.

In the following, we provide details and outline a Gibbs sampling algorithm to estimate this predictive process model.

2.1 Model implementation

Let $\mathcal{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n\}$ be the number of locations that have yielded observations on the outcome and the predictors. For now, assume perfect space-time alignment—

in other words, every time point t has yielded observations over all the locations in $\mathcal{S}$ and there are no empty cells in the space-time data matrix.

Let $\boldsymbol{y}_t = (y_t(\boldsymbol{s}_1), y_t(\boldsymbol{s}_2), \ldots, y_t(\boldsymbol{s}_n))'$ denote the $n \times 1$ vector of outcomes and $\boldsymbol{X}_t$ be the $n \times p$ matrix whose i-th row is given by $\boldsymbol{x}_t(\boldsymbol{s}_i)'$. The model yielding the likelihood can now be written down as

$$\boldsymbol{y}_t = \boldsymbol{X}_t\boldsymbol{\beta}_t + \tilde{\boldsymbol{u}}_t + \boldsymbol{\epsilon}_t; \quad \boldsymbol{\epsilon}_t \overset{\text{ind.}}{\sim} N(\boldsymbol{0}, \tau_t^2\boldsymbol{I}_n), \quad t = 1, 2, \ldots, N_t \ . \tag{2}$$

Here $\tilde{\boldsymbol{u}}_t = (\tilde{u}_t(\boldsymbol{s}_1), \tilde{u}_t(\boldsymbol{s}_2), \ldots, \tilde{u}_t(\boldsymbol{s}_n))'$ is the realization of the spatial predictive process. In particular, this implies that

$$\tilde{\boldsymbol{u}}_t = \sum_{k=1}^{t}\left[\boldsymbol{C}_k(\boldsymbol{\theta}_k)'\boldsymbol{C}_k^*(\boldsymbol{\theta}_k)^{-1}\boldsymbol{w}_k^* + \tilde{\boldsymbol{\epsilon}}_k\right], \quad t = 1, 2, \ldots, N_t \ ,$$

where each $\boldsymbol{C}_k(\boldsymbol{\theta}_k)'$ denotes the $n \times n^*$ matrix whose (i, j)-th element is the covariance between $w_k(\boldsymbol{s}_i)$ and $w_k(\boldsymbol{s}_j{}^*)$ given by $C_k(\boldsymbol{s}_i, \boldsymbol{s}_j^*; \boldsymbol{\theta}_k)$ and $\boldsymbol{C}_k^*(\boldsymbol{\theta}_k)$ is the $n^* \times n^*$ matrix with (i, j)-th element $C_k(\boldsymbol{s}_i{}^*, \boldsymbol{s}_j{}^*)$. Also, $\tilde{\boldsymbol{\epsilon}}_k \overset{\text{ind.}}{\sim} N(\boldsymbol{0}, \tilde{\boldsymbol{D}}_k)$, where $\tilde{\boldsymbol{D}}_k$ is an $n \times n$ diagonal matrix with $C_k(\boldsymbol{s}_i, \boldsymbol{s}_i) - \boldsymbol{c}_k(\boldsymbol{s}_i)'\boldsymbol{C}_k^*(\boldsymbol{\theta})^{-1}\boldsymbol{c}_k(\boldsymbol{s}_i)$ as the i-th diagonal element, where $\boldsymbol{c}_k(\boldsymbol{s}_i)'$ is the i-th row of $\boldsymbol{C}_k(\boldsymbol{\theta}_k)'$. Bayesian inference seeks the posterior distribution of the parameters $\boldsymbol{\beta}_0$, $\boldsymbol{\Sigma}_\eta$, $\{\boldsymbol{\theta}_t\}$, $\{\boldsymbol{\beta}_t\}$, $\{\tilde{\boldsymbol{u}}_t\}$, $\{\boldsymbol{w}_t^*\}$ and $\{\tau_t^2\}$. The full hierarchical model can be written down as:

$$\begin{aligned}
&\prod_{t=1}^{N_t} p(\boldsymbol{\theta}_t) \prod_{t=1}^{N_t} \text{IG}\left(\tau_t^2 | a_\tau, b_\tau\right) N(\boldsymbol{\beta}_0 | \boldsymbol{m}_0, \boldsymbol{\Sigma}_0) \text{IW}(\boldsymbol{\Sigma}_\eta | r_\eta, \Upsilon_\eta) \prod_{t=1}^{N_t} N(\boldsymbol{\beta}_t | \boldsymbol{\beta}_{t-1}, \boldsymbol{\Sigma}_\eta) \\
&\prod_{t=1}^{N_t} N(\boldsymbol{w}_t^* | \boldsymbol{0}, \boldsymbol{C}_t^*(\boldsymbol{\theta}_t)) \prod_{t=1}^{N_t} N(\tilde{\boldsymbol{u}}_t | \tilde{\boldsymbol{u}}_{t-1} + \boldsymbol{C}_t(\boldsymbol{\theta}_t)'\boldsymbol{C}_t^*(\boldsymbol{\theta}_t)^{-1}\boldsymbol{w}_t^*, \tilde{\boldsymbol{D}}_t) \\
&\prod_{t=1}^{N_t} N(\boldsymbol{y}_t | \boldsymbol{X}_t\boldsymbol{\beta}_t + \tilde{\boldsymbol{u}}_t, \tau_t^2\boldsymbol{I}_n) \ ,
\end{aligned} \tag{3}$$

with notations as in Gelman et al. (2004). Our proposed Gibbs sampler algorithm is detailed in Appendix 1.

This sampler easily adapts to situations when $y_t(\boldsymbol{s}_i)$ is missing (or not monitored) for some $\boldsymbol{s}_i$'s at certain time points. We simply treat such variables as missing values and update them, in each iteration, from their associated full conditional distribution, which is $N(\boldsymbol{x}_t(\boldsymbol{s}_i)'\boldsymbol{\beta}_t + \tilde{u}_t(\boldsymbol{s}_i), \tau_t^2)$. We assume that all predictors in $\boldsymbol{x}_t(\boldsymbol{s}_i)$ will be available in the space-time data matrix, so this spatial interpolation step for missing outcomes is straightforward and inexpensive. This is exactly how we also perform spatial prediction at arbitrary locations within the domain.

2.2 Model assessment

We assess model performance and subsequent comparisons by simulating *independent* replicates for each observed outcome: for each time point t, we sample from the distribution $p(y_{rep,t}(\boldsymbol{s}_i)|\boldsymbol{y}) = \int N(y_{\text{rep},t}(\boldsymbol{s}_i)|\boldsymbol{x}_t(\boldsymbol{s}_i)'\boldsymbol{\beta}_t + \tilde{u}_t(\boldsymbol{s}_i), \tau_t^2)p(\boldsymbol{\beta}_t, \tilde{u}_t(\boldsymbol{s}_i), \tau_t^2|\boldsymbol{y})$

where $\boldsymbol{y}$ is the collection of all observed outcomes and $p(\boldsymbol{\beta}_t, \tilde{u}_t(\boldsymbol{s}_i), \tau_t^2|\boldsymbol{y})$ is the marginal posterior distribution of the unknowns in the likelihood. Sampling from the posterior predictive distribution is straightforward using *composition*: for each sampled value of $\{\boldsymbol{\beta}_t, \tilde{u}_t(\boldsymbol{s}_i), \tau_t^2\}$ from the marginal posterior, we draw y_{rep}, $t(\boldsymbol{s}_i)$, one for one, from $N(\boldsymbol{x}_t(\boldsymbol{s}_i)'\boldsymbol{\beta}_t + \tilde{u}_t(\boldsymbol{s}_i), \tau_t^2)$.

We will prefer models that will perform well under a decision-theoretic balanced loss function, penalizing both departure of replicated means from their observed values (lack of fit) and excessive uncertainty in the replicated data (predictive variance). Using a squared error loss function (e.g., Gelfand and Ghosh 1998), the measures for these two criteria are evaluated as $G = \sum_{t=1}^{N_t} \sum_{i=1}^{n} (y_t(\boldsymbol{s}_i) - \mu_{\text{rep},t}(\boldsymbol{s}_i))^2$ and $P = \sum_{t=1}^{N_t} \sum_{i=1}^{n} \sigma^2_{\text{rep},t}(\boldsymbol{s}_i)$, respectively, where $\mu_{\text{rep},t}(\boldsymbol{s}_i) = E[y_{\text{rep},t}(\boldsymbol{s}_i) \mid \boldsymbol{y}]$ and $\sigma^2_{\text{rep},t}(\boldsymbol{s}_i) = \text{var}[y_{\text{rep},t}(\boldsymbol{s}_i)| \boldsymbol{y}]$ are the posterior predictive mean and variance for each $y_{\text{rep},t}(\boldsymbol{s}_i)$. We will use the score $D = G + P$ as a model selection criterion, with lower values of D indicating better models.

3 Weather station analysis

A host of research questions in climate sciences require estimates of weather variables at points in space and time where observed data are not available. For instance, in an effort to forecast regional climates, researchers are exploring methods to validate and quantify uncertainty in regional climate model (RCM) outputs, such as those produced as part of the North American Regional Climate Change Assessment Program (Meehl et al. 2007). These efforts often entail coupling RCM hindcast outputs with temporally and spatially coinciding historic weather records (see, e.g., Salazar et al. 2011; Tebaldi and Sansó 2008; Smith et al. 2009; Rougier et al. 2010). A challenge in these model exercises is to rectify the pervasive issue of spatial and temporal misalignment between the RCM output pixels and sparse point-referenced weather station data. A solution is to use the sparse weather station data to predict the weather variables of interest at the RCM grid locations and time intervals. This missingness and subsequent need to complete these records motivates the following illustrative analysis.

The spatial domain shown in Fig. 1a contains 1,530 weather stations equipped with sensors that recorded temperature between 1/1/2000 and 1/1/2005. Using these data, the outcome variable $y_t(\boldsymbol{s})$ is mean monthly temperature in station location $\boldsymbol{s}$ at time t. These data were retrieved from the National Climatic Data Center's (NCDC) web-based interface (http://lwf.ncdc.noaa.gov/oa/climate/climatedata.html) and converted from degrees Fahrenheit to Celsius. Geographic coordinates for the stations were projected from decimal degrees to a Transverse Mercator projection to provide comfort with Euclidean distance metrics. Fig. 1a illustrates the study area and station locations as well as an elevation surface, which was used as a predictor in the subsequent analysis. Due to sensor or data management failure, $\sim$16% of the monthly station records in our dataset were missing (i.e., of the $n \times N_t$ = 1,530 station $\times$ 61 months = 93,330 possible observations, 15,107 had missing outcomes). Fig. 2 shows the distribution of missing observations within the dataset.

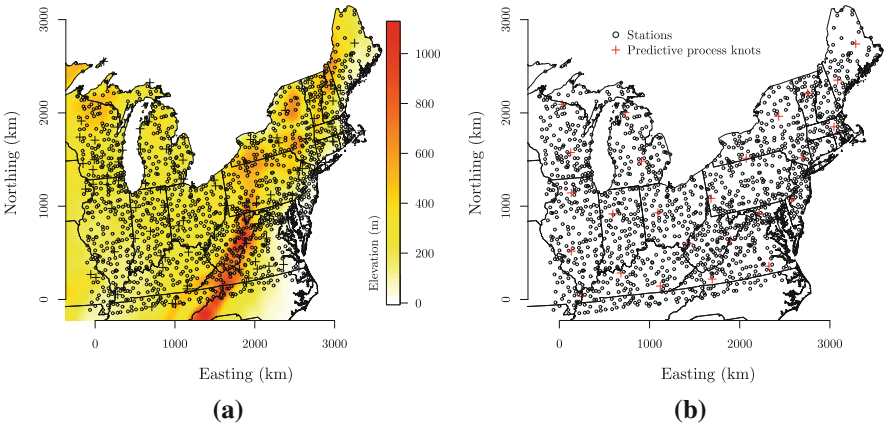

Fig. 1 **a** Observed and model validation weather station locations and elevation surface. **b** Locations of predictive process knots for the 25 knot model with weather station locations

The data were analyzed using five candidate submodels of (2), including the non-spatial (i.e., $\tilde{\boldsymbol{u}}_t = \mathbf{0}$ for $t = 1, 2, \ldots, N_t$) and spatial predictive process models based on 5, 10, 25, and 50 knot intensities. For all models, an intercept and station elevation form the 1×2 $\boldsymbol{x}_t(\boldsymbol{s})'$ vector corresponding to $y_t(\boldsymbol{s})$. For the spatial models, knots were placed at the centroid of station location clusters, which were delineated using a *k-means* clustering algorithm (Hartigan and Wong 1979). Knot locations for the 25 knot model are illustrated in Fig. 1b. Performance of the candidate models was evaluated using the G, P, and D criterion detailed in Sect. 2.2 and root mean squared prediction error (RMSPE) calculated using a 10% hold-out set of 10 randomly selected months within each of 100 randomly selected stations. Due to actual missingness, RMSPE was based on only 823 of the 1,000 selected hold-out months. Station time series that contain hold-out months are indicated in Fig. 1a with a '+' symbol.

To complete the model specification, we assigned vague hyperparameters $\boldsymbol{\Sigma}_0 = 1,000\boldsymbol{I}_2$ and $\boldsymbol{m}_0 = (0, 0)$ for the normal prior on $\boldsymbol{\beta}_0$, inverse-Wishart IW$(2, 0.01\boldsymbol{I}_2)$ prior on the $2 \times 2\,\boldsymbol{\Sigma}_\eta$, and inverse-Gamma IG$(2, \cdot)$ for the τ_t^2's and σ_t^2's. For the non-spatial model, the IG scale hyperparameter was set to 10, which is approximately the average residual variance estimated using a simple time invariant linear regression within each of the N_t months. For the spatial models, the IG scale for τ_t^2's and σ_t^2's was set to 5. Given a maximum distance of 2,305 km between any two stations and assuming an exponential spatial correlation function, we set the support for the spatial decay parameters ϕ_t's to follow a Uniform $U(0.001, 0.03)$, which corresponds to 100 to 3,000 km.

The sampler detailed in Appendix 1 was implemented in `R` (http://www.r-project.org), which was compiled with Intel's Math Kernel Library threaded `BLAS` and `LAPACK` routines. All analyses were conducted on a Linux workstation using an Intel Nehalem quad-Xeon processor. Three MCMC chains were run for 15,000 iterations each. Convergence was diagnosed using the `CODA` package in `R`, by monitoring mixing of chains and the Gelman-Rubin statistic (Gelman and Rubin

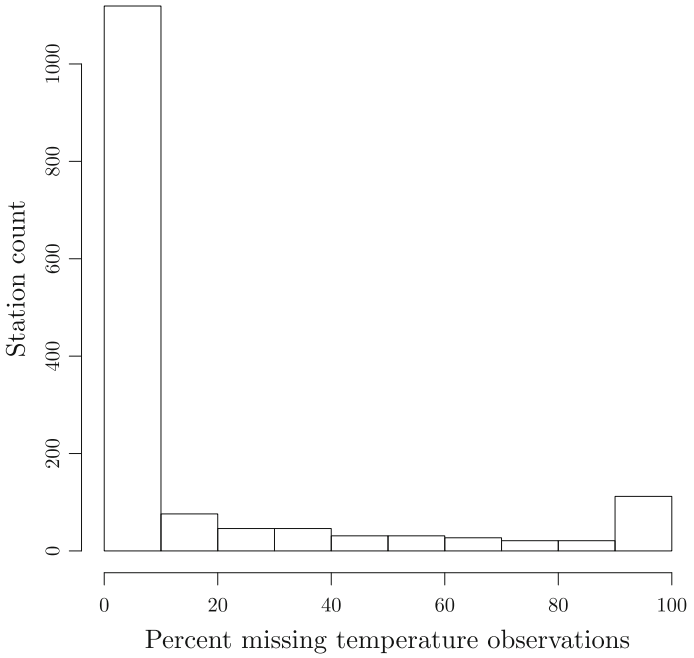

Fig. 2 Frequency of missing monthly temperature observations within the study dataset

1992). Satisfactory convergence was diagnosed within 5,000 iterations for all models and posterior inference was based on a post-burn-in sample of 15,000 iterations (5,000 from each chain).

The first three rows of Table 1 summarize the fit of the candidate models. Here, compared to the non-spatial model, the lower G and D values for the predictive process models suggest the addition of the structured $\tilde{\boldsymbol{u}}_t$ substantially improves model fit. The non-spatial model's D value of 1,398,210 is substantially larger than that of the spatial model and reflects the ability of the spatio-temporal random effects to enable closer approximation to the observed data. Fig. 3 illustrates a scatterplot of μ_{rep} versus the observed temperature at each station for the non-spatial and 25-knot predictive process model ((a) and (b) respectively). The much improved goodness of fit over the vast number of observations results in the significantly smaller G for the spatio-temporal models. The disparity between the replicated

Table 1 Candidate model comparison using G, P, D, and holdout set root mean squared prediction error (RMSPE)

	Non-spatial	Predictive process knots			
		5	10	25	50
G	699,174	282	162	75	80
P	699,036	8,333	6,829	6,022	6,030
D	1,398,210	8,616	6,992	6,097	6,110
RMSPE	8.4	0.54	0.45	0.31	0.29

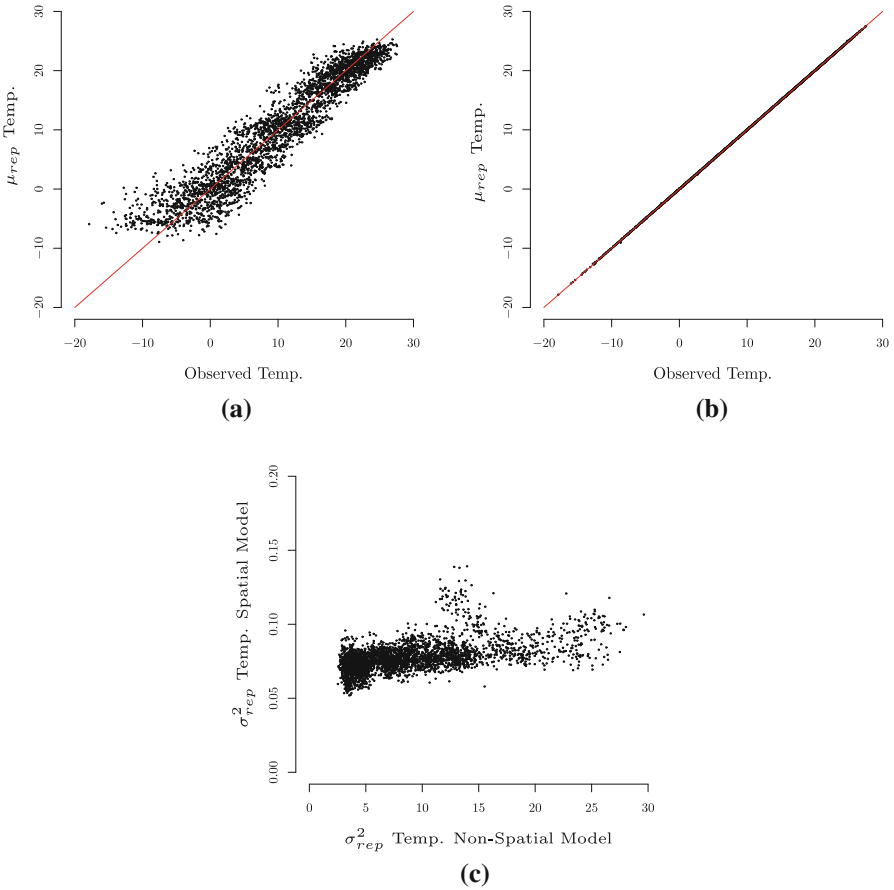

Fig. 3 Observed temperature versus μ_{rep} for the **a** non-spatial and **b** 25 knot predictive process models. **c** Non-spatial and 25 knot predictive process models σ^2_{rep}

precision σ^2_{rep} of the non-spatial and predictive process models is illustrated in Fig. 3c (notice the difference in scales). These contribute to the substantial reduction in D. Returning to Table 1, among the spatial models, increasing the knot intensity beyond 25 knots does not seem to improve model fit, at least based on the G and D criteria.

There is negligible difference between the parameter estimates and predictive inference between the 25 and 50 knot model. Therefore, for brevity and because the 25-knot model produces a D that is marginally lower than the 50 knot model, the subsequent results focus on the 25 knot model.

Figures 4 and 5 display $\beta_t = (\beta_{0,t}, \beta_{Elev,t})$, the time-varying intercept and elevation slope parameters for the non-spatial and 25 knot predictive process model, respectively. As expected, for both models, seasonal temperature oscillation is clearly seen in these estimates—the intercept is lower in winter and higher in summer. In this way, a seasonal pattern is retrieved although no such structure was

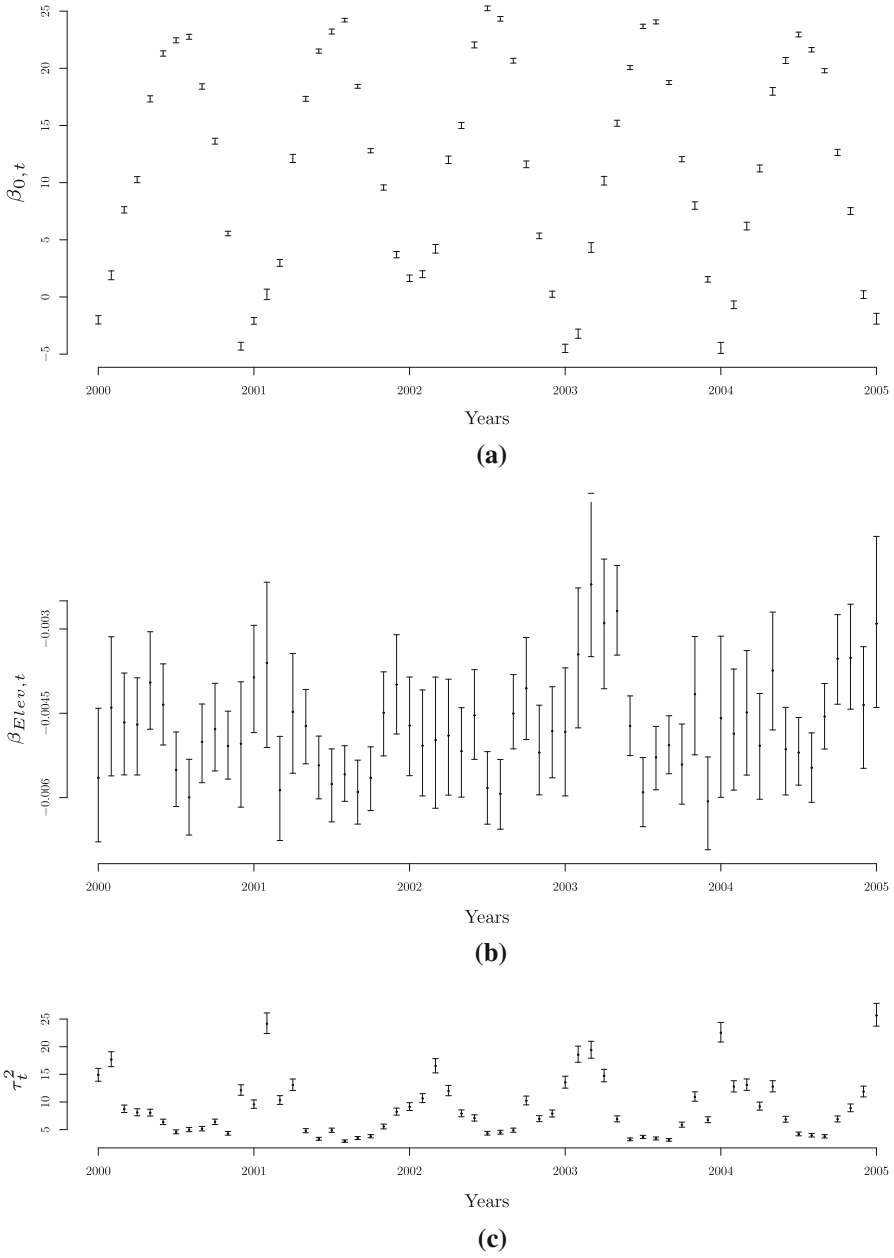

Fig. 4 Parameter credible intervals, 50 (2.5 97.5) percentiles, non-spatial model **a** $\beta_{0,t}$, **b** $\beta_{Elev,t}$, and **c** τ_t^2

imposed. Here, too, we see that the negative impact of elevation on temperature is significant across all months.

The candidate models produced nearly identical estimates of Σ_η. Table 2 offers these estimates for the non-spatial and 25 knot predictive process models. Here the

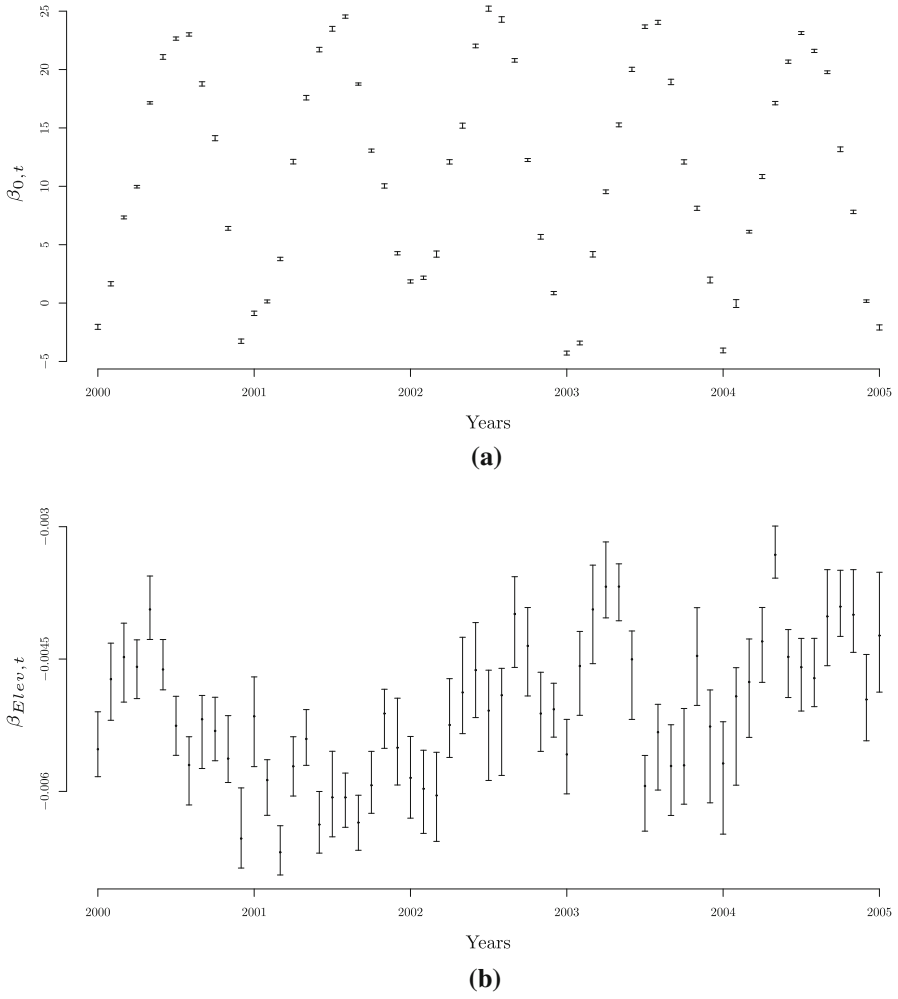

Fig. 5 Slope parameter credible intervals, 50 (2.5 97.5) percentiles, for the 25 knot predictive process model **a** $\beta_{0,t}$, and **b** $\beta_{Elev,t}$

relatively large value of $\boldsymbol{\Sigma}_{\eta}[1,1]$ reflects the large seasonal fluctuation among the $\beta_{0,t}$'s.

Figures 4c, 6a and b show the time-varying variance parameters for the non-spatial and 25 knot predictive process models. Comparing among these figures, it is clear the spatial model reapportions the majority of the residual variance to the spatio-temporal random effects. Again, it is this difference between the models that results in the disparity in model fit, D, and associated dispersion of the replicated data in Fig. 3.

The model's ability to fit local temperature trends, and hence predictive ability, is partially determined by the minimum distance among the predictive process knots. This distance influences the estimates of ϕ_t and σ_t^2—smaller distances among the

Table 2 Σ_η parameter credible intervals, 50 (2.5 97.5) percentiles, for the non-spatial and 25 knot predictive process models

	Non-spatial	Predictive process knots 25 knot
$\Sigma_\eta[1,1]$	24.41 (17.97, 36.65)	23.61 (16.83, 35.45)
$\Sigma_\eta[2,1]$	0.001 (−0.050, 0.054)	0.001 (−0.050, 0.055)
$\Sigma_\eta[2,2]$	0.002 (0.001, 0.002)	0.002 (0.001, 0.002)

knots will allow for detection of local or short-range trends in the latent spatio-temporal process. Obviously, the more knots the better the approximation of the process; however, the choice of knot intensity is determined by computational resources. Here, the 5, 10, 25, and 50 knot models had minimum distances of 474, 242, 179, and 117 km, and corresponding run-times of approximately 8, 9, 14, and 20 hours, respectively.

The effective spatial ranges (the distance at which the exponential correlation function equals 0.05 i.e., $-\log(0.05)/\phi_t$) also show some seasonal trends. Specifically, Fig. 6c suggests the spatial range increases in the winter and decreases in the summer. The other predictive process models produced similar trends.

The last row in Table 1 provides RMSPE calculated using the observed temperature and the median of the candidate models' posterior predictive distribution at 823 hold-out months. As suggested by the values of D, the RMSPE shows that the addition of the random effects greatly improves prediction. Further, increasing knot intensity allows the models to more accurately conform to local temperature patterns and hence provide better prediction—reflected in a decrease in RMSPE as the number of knots increases.

Figure 7 illustrates the temperature time series for four of the 100 stations that contain hold-out months used to calculate the RMSPE in Table 1. Within each plot, the black solid line is the observed temperature. The absence of the black solid line in portions of Fig. 7a–c means the temperature record is missing due to sensor failure or a data reporting issue for those months. The '○' symbol along the x-axis of each series indicates hold-out months (i.e., the corresponding temperature value was purposefully removed and did not contribute to model parameter estimation). The gray region and fine dotted line represent the 95% credible interval and median, respectively, of the posterior predictive distribution of the non-spatial model. Similarly, the three long dashed lines correspond to the 25 knot model's 2.5, 50, and 97.5 posterior predictive distribution percentiles. Clearly, the broadness of the non-spatial model's posterior predictive distribution is a result of its τ_t^2, which is large relative to that of the spatial model. The non-spatial model's 95% credible interval covers the *true* temperature for 96% of the hold-out months, where as the spatial model's coverage is $\sim$97% across knot intensities. As expected, there is a perceptible widening of the spatial model's credible interval for months with missing data. For example, in Fig. 7a the interval widens for the first three hold-out months in early 2001 and is substantially wider in 2004 to reflect the missing data.

Appendix 2 (Electronic supplementary material) offers a dynamic plot of the observed and predicted temperatures interpolated over the weather station locations for each of the 61 months. Here, the top row illustrates the observed station

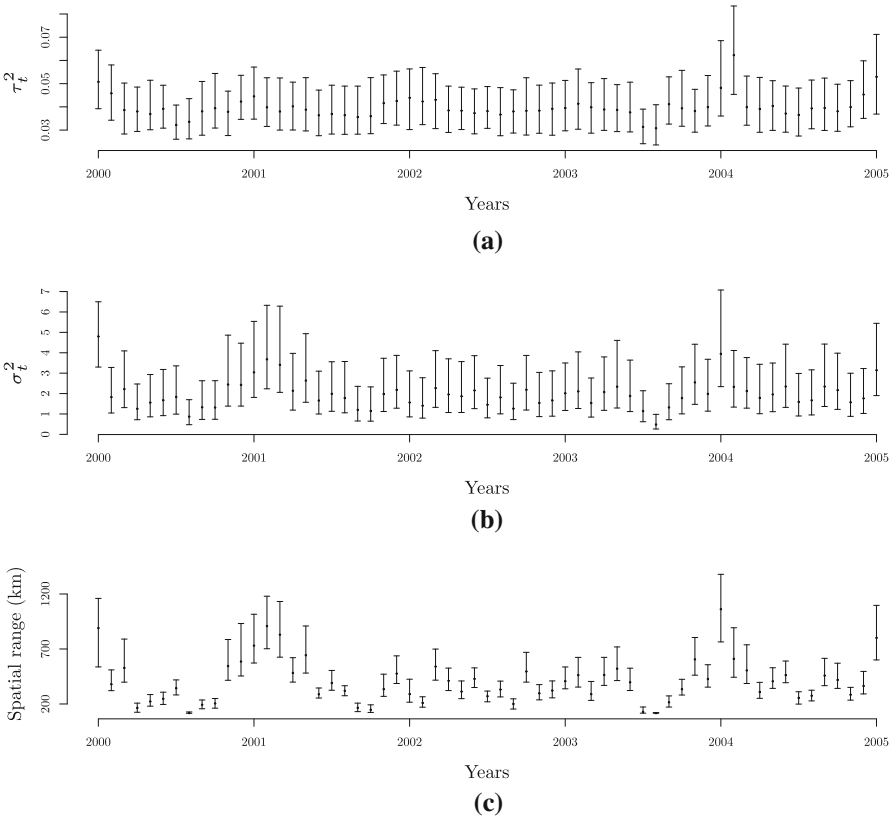

Fig. 6 Residual parameter credible intervals, 50 (2.5 97.5) percentiles, for the 25 knot predictive process model **a** τ_t^2, **b** σ_t^2, and **c** effective spatial range (i.e., $-\log(0.05)/\phi_t$)

temperature (left) and 50 knot predictive process model's median posterior predictive distribution (right). As suggested by the model fit assessment and Fig. 7, the predictive process model closely approximates the spatial and temporal temperature dynamics. The second row decomposes the 50 knot predictive process model's prediction into the median of the posterior predicted $\boldsymbol{x}_t(\boldsymbol{s})'\boldsymbol{\beta}_t$ (left) and $\tilde{\boldsymbol{u}}_t$ (right), which allows us to more easily see the contribution of the mean trend (i.e., predictors) and the spatio-temporal random effect.

4 Closing remarks

We have proposed and illustrated one approach to addressing the problem of fitting the flexible, but computationally demanding, dynamic modeling framework for spatio-temporal data proposed by Gelfand et al. (2005) to large datasets. To do so, we simply replaced the parent spatial process by its induced predictive process. The appeal of this solution is that one need not digress from the modeling objectives to

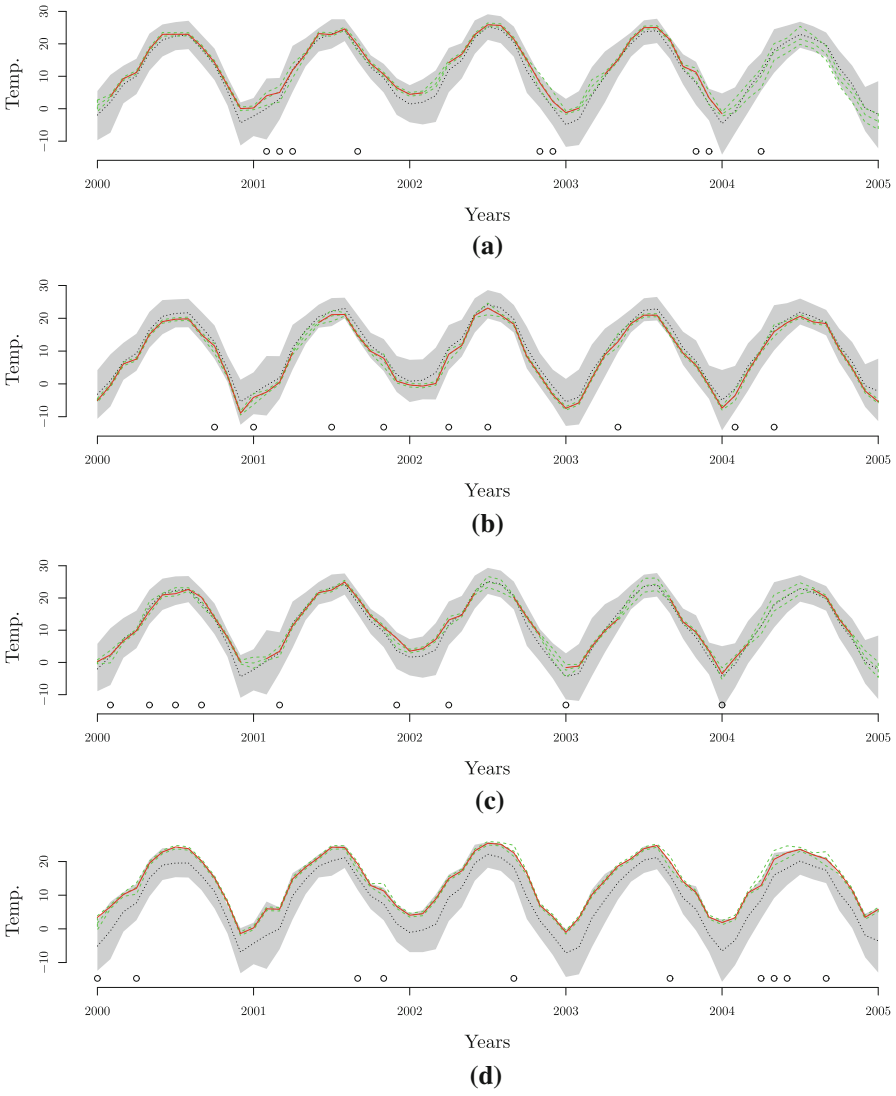

Fig. 7 **a–d** Four of 100 hold-out locations where temperature predictions were made. *Open circle symbols* indicate those months that were used for prediction validation. *The solid line* is the observed temperature. *Gray region* is the non-spatial model's 95% credible interval, and the fine *dotted line* is the associated posterior predicted median. *Long dashed lines* indicate the 25 knot predictive process model's 95% credible interval and associated posterior predictive median

think about choices of basis functions, or kernels or alignment algorithms for the locations. Note that even if the parent process is stationary, the predictive processes derived from it are nonstationary. This makes them more flexible and adaptable for explaining datasets that show departures from stationarity (stationarity is rarely seen in practice) and lead to improved model performance.

For clarity of illustration, we have focused only upon the spatially varying intercept model with a single outcome. We recognize that this is a simplification from Gelfand et al. (2005), who used a vector of spatial processes for all the regression coefficients. The predictive process counterpart of these, more general, models would simply replace the vector of spatial processes with a joint or multivariate predictive process; see, e.g., Banerjee et al. (2008). Another simplifying assumption we make is that of a random walk for the $\boldsymbol{\beta}_t$. In standard dynamic linear models (West and Harrison, 1997), the so-called *system matrices* that appear in the measurement and transition equations may change over time and can depend upon parameters. The system matrix in the measurement equation is customarily specified from design considerations for the problem at hand (as we have done here using spatially referenced predictors), while that in the transition equation is specified using modeling assumptions. We could generalize the transition equation in (1) using $\boldsymbol{\beta}_t = \boldsymbol{G}_t\boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t$, where $\boldsymbol{G}_t$ is a $p \times p$ transition system matrix, and proceed with the corresponding predictive process model in the usual way. However, taking $\boldsymbol{G}_t$ to be the identity matrix produces a nonstationary multivariate AR(1) model over time, which is simple to interpret and adequately flexible for most environmental datasets in practice.

As in other low-rank kriging approaches, knot selection is required and the number of knots will impact model fit and prediction, although, in our experience, for most applications a reasonable grid of knots should lead to robust inference. However, we note that with fewer knots the separation between them increases and estimating the latent process with fine scale spatial dependence becomes difficult.

The weather station temperature analysis demonstrated that a low-dimensional representation of the parent process is highly effective at borrowing information over space and time to deliver accurate and precise prediction for missing sensor readings and new locations. Further, by working in a Bayesian paradigm we have access to the full posterior predictive distribution at each new location and time. This could be useful in studying exceedance probabilities for temperatures. In addition to the climate science modeling scenario that opens Sect. 3, this modeling framework could also be used in settings where the investigator wishes to propagate uncertainty in weather variables through subsequent economic or ecological numerical models. For example, many agriculture crop yield models require spatially and temporally explicit temperature and precipitation inputs which must be somehow imputed to the region or field of interest.

Finally, we recognize that the `R` implementation of the proposed sampler is very inefficient. The sampler requires many nested loops over the spatial and temporal dimensions. `R`, like most other interpreted programming languages, is notoriously slow at executing *for* loops. This computational hurdle can be all but eliminated by implementing the code in C/C++ or some other compiled language, which we plan to do and make freely available to encourage further exploration and comparison.

Acknowledgments The work of the first author was supported by the USDA Forest Inventory and Analysis National Program and Forest Health Technology Enterprise Team. The first two authors were supported by the National Science Foundation (NSF) grant NSF-DMS-1106609 and National Institutes of Health (NIH) grant NIH/NIGMS 1-RC1-GM092400-01. The third author was supported by NSF-CDI-0940671 and NSF-DMS-0914906.

Appendix 1: A Gibbs sampling algorithm for Bayesian inference

The Gibbs sampler algorithm for (3) is initiated by set all parameters to some starting values. Then, in each iteration of the Gibbs sampler we will update:

1. $\boldsymbol{\beta}_0 \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\beta}_0|\cdot}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_0|\cdot}\right)$, where

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}_0|\cdot} = \left(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_\eta^{-1}\right)^{-1}$$
$$\boldsymbol{\mu}_{\boldsymbol{\beta}_0|\cdot} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_0|\cdot}\left[\boldsymbol{\Sigma}_0^{-1}\boldsymbol{m}_0 + \boldsymbol{\Sigma}_\eta^{-1}\boldsymbol{\beta}_1\right]$$

2. For each t = 1, 2, …, N_t, update the following:

   (a) $\boldsymbol{\beta}_t \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\beta}_t|\cdot}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t|\cdot}\right)$, where

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t|\cdot} = \left(2\boldsymbol{\Sigma}_\eta^{-1} + \frac{1}{\tau_t^2}\boldsymbol{X}_t'\boldsymbol{X}_t\right)^{-1}$$
$$\boldsymbol{\mu}_{\boldsymbol{\beta}_t|\cdot} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t|\cdot}\left[\boldsymbol{\Sigma}_\eta^{-1}(\boldsymbol{\beta}_{t-1} + \boldsymbol{\beta}_{t+1}) + \boldsymbol{X}_t'\frac{(\boldsymbol{y}_t - \tilde{\boldsymbol{u}}_t)}{\tau_t^2}\right].$$

   For $t = N_t$, we set $\boldsymbol{\beta}_{t+1} = \boldsymbol{0}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t|\cdot} = \left(\boldsymbol{\Sigma}_\eta^{-1} + \frac{1}{\tau_t^2}\boldsymbol{X}_t'\boldsymbol{X}_t\right)^{-1}$.

   (b) Set $\boldsymbol{F}_t(\boldsymbol{\theta}_t) = \boldsymbol{C}_t(\boldsymbol{\theta}_t)'\boldsymbol{C}_t^*(\boldsymbol{\theta}_t)^{-1}$. Then, update $\boldsymbol{w}_t^* \sim N\left(\boldsymbol{\mu}_{\boldsymbol{w}_t^*|\cdot}, \boldsymbol{\Sigma}_{\boldsymbol{w}_t^*|\cdot}\right)$, where

$$\boldsymbol{\Sigma}_{\boldsymbol{w}_t^*|\cdot} = \left(\boldsymbol{C}_t^*(\boldsymbol{\theta}_t)^{-1} + \boldsymbol{F}_t(\boldsymbol{\theta}_t)'\tilde{\boldsymbol{D}}_t^{-1}\boldsymbol{F}_t(\boldsymbol{\theta}_t)\right)^{-1}$$
$$\boldsymbol{\mu}_{\boldsymbol{w}_t^*|\cdot} = \boldsymbol{\Sigma}_{\boldsymbol{w}_t^*|\cdot}\boldsymbol{F}_t(\boldsymbol{\theta}_t)'\tilde{\boldsymbol{D}}_t^{-1}(\tilde{\boldsymbol{u}}_t - \tilde{\boldsymbol{u}}_{t-1})$$

   (c) $\tilde{\boldsymbol{u}}_t \sim N\left(\boldsymbol{\mu}_{\tilde{\boldsymbol{u}}_t|\cdot}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{u}}_t|\cdot}\right)$, where

$$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{u}}_t|\cdot} = \left(\tilde{\boldsymbol{D}}_t^{-1} + \tilde{\boldsymbol{D}}_{t+1}^{-1} + \frac{1}{\tau_t^2}\boldsymbol{I}_n\right)^{-1}$$
$$\boldsymbol{\mu}_{\tilde{\boldsymbol{u}}_t|\cdot} = \boldsymbol{\Sigma}_{\tilde{\boldsymbol{u}}_t|\cdot}\left[\tilde{\boldsymbol{D}}_t^{-1}(\tilde{\boldsymbol{u}}_{t-1} + \boldsymbol{F}_t(\boldsymbol{\theta}_t)\boldsymbol{w}_t^*) + \tilde{\boldsymbol{D}}_{t+1}^{-1}(\tilde{\boldsymbol{u}}_{t+1} - \boldsymbol{F}(\boldsymbol{\theta}_{t+1})\boldsymbol{w}_{t+1}^*) + \frac{\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta}_t}{\tau_t^2}\right].$$

   For $t = N_t$, we set $\tilde{\boldsymbol{D}}_{t+1}^{-1} = \boldsymbol{0}$.

   (d) $\tau_t^2 \sim \mathrm{IG}(a_{\tau|\cdot}, b_{\tau|\cdot})$, where

$$a_{\tau|\cdot} = a_\tau + \frac{n}{2} \quad \text{and} \quad b_{\tau|\cdot} = b_\tau + \frac{1}{2}(\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta}_t - \tilde{\boldsymbol{u}}_t)'(\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta}_t - \tilde{\boldsymbol{u}}_t)$$

   (e) Note that $\boldsymbol{\theta}_t = \{\sigma_t^2, \theta_{1t}, \theta_{2t}\}$. Let the prior on these parameters be of the form $\mathrm{IG}(\sigma_t^2 \mid a_\sigma, b_\sigma)\ p(\theta_{1t}, \theta_{2t})$. Then, we can update $\sigma_t^2 \sim \mathrm{IG}(a_{\sigma|\cdot}, b_{\sigma|\cdot})$, where

$$a_{\sigma|\cdot} = a_\sigma + \frac{n^*}{2} \quad \text{and} \quad b_{\sigma|\cdot} = b_\sigma + \frac{1}{2}\boldsymbol{w}_t^{*\prime}\boldsymbol{R}_t^*(\boldsymbol{\theta}_t)^{-1}\boldsymbol{w}_t^* \ ,$$

where $\boldsymbol{C}_t^*(\boldsymbol{\theta}_t) = \sigma_t^2\boldsymbol{R}_t^*(\boldsymbol{\theta}_t)$. The remaining members of $\boldsymbol{\theta}_t$ are updated using random-walk Metropolis steps. The target proposal is:

$$p(\boldsymbol{\theta}_t)N(\boldsymbol{w}_t^*|\boldsymbol{\theta}, \boldsymbol{C}_t^*(\boldsymbol{\theta}_t))N(\tilde{\boldsymbol{u}}_t|\tilde{\boldsymbol{u}}_{t-1} + \boldsymbol{F}_t(\boldsymbol{\theta}_t)\boldsymbol{w}_t^*) \ ,$$

where $\boldsymbol{F}_t(\boldsymbol{\theta}_t)$ is as defined in Step (b).

3. Assuming that $\boldsymbol{\Sigma}_\eta \sim \text{IW}(r_\eta, \Upsilon_\eta)$ is the Inverse-Wishart prior density parametrized as:

$$\text{IW}(\boldsymbol{\Sigma}_\eta|r_\eta, \Upsilon_\eta) \propto \frac{1}{|\boldsymbol{\Sigma}_\eta|^{r_\eta+(p+1)/2}}\exp\left\{-\frac{1}{2}tr(\Upsilon_\eta^{-1}\boldsymbol{\Sigma}_\eta^{-1})\right\},$$

$\boldsymbol{\Sigma}_\eta \sim \text{IW}\left(r_{\eta|\cdot}, \Upsilon_{\eta|\cdot}\right)$, where

$$r_{\eta|\cdot} = r_\eta + N_t \quad \text{and} \quad \Upsilon_{\eta|\cdot} = \left[\Upsilon_\eta^{-1} + \sum_{i=1}^{N_t}\sum_{j=1}^{N_t} a^{ij}\left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_0\right)\left(\boldsymbol{\beta}_i - \boldsymbol{\beta}_0\right)'\right]^{-1} ,$$

where a^{ij} is the (i, j)-th element of the matrix $\boldsymbol{A}^{-1}$, and

$$\boldsymbol{A}^{-1} = \begin{pmatrix} 1 & -1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{pmatrix} .$$

Alternatively, we can also update $\boldsymbol{\Sigma}_\eta$ from $\text{IW}\left(r_{\eta|\cdot}, \Upsilon_{\eta|\cdot}\right)$, where

$$r_{\eta|\cdot} = r_\eta + N_t \quad \text{and} \quad \Upsilon_{\eta|\cdot} = \left[\Upsilon_\eta^{-1} + \sum_{t=1}^{N_t}(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})'\right]^{-1} .$$

References

Banerjee S, Carlin BP, Gelfand AE (2004) Hierarchical modeling and analysis for spatial data. Chapman and Hall/CRC Press, Boca Raton

Banerjee S, Gelfand AE, Finley AO, Sang H (2008) Gaussian predictive process models for large spatial datasets. J R Stat Soc Ser B 70(4):825–848

Carlin BP, Louis TA (2000) Bayes and empirical Bayes methods for data analysis, 2nd edn. Chapman and Hall/CRC Press, Boca Raton [FL]

Chilés JP, Delfiner P (1999) Geostatistics: modelling spatial uncertainty. Wiley, New York

Crainiceanu CM, Diggle PJ, Rowlingson B (2008) Bivariate binomial spatial modeling of Loa loa Prevalence in tropical Africa (with discussion). J Am Stat Assoc 103(481):21–37

Cressie NAC (1993) Statistics for spatial data, 2nd edn. Wiley, New York

Cressie NAC, Johannesson G (2008) Fixed rank kriging for very large spatial data sets. J R Stat Soc Ser B 70(1):209–226

Du J, Zhang H, Mandrekarm VS (2009) Fixed-domain asymptotic properties of tapered maximum likelihood estimators. Ann Stat 37(6A):3330–3361
Finley AO, Sang H, Banerjee S, Gelfand AE (2009) Improving the performance of predictive process modeling for large datasets. Comput Stat Data Anal 53(8):2873–2884
Fuentes M (2007) Approximate likelihood for large irregularly spaced spatial data. J Am Stat Assoc 102(477):321–331
Furrer R, Genton MG, Nychka D (2006) Covariance tapering for interpolation of large spatial datasets. J Comput Graph Stat 15(3):502–523
Gelfand AE, Banerjee S, Gamerman D (2005) Univariate and multivariate dynamic spatial modelling. Environmetrics 16(5):465–479
Gelfand AE, Ghosh SK (1998) Model Choice: a minimum posterior predictive loss approach. Biometrika 85(1):1–11
Gelman A, Rubin D (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7(4):457–511
Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis, 2nd edn. Chapman and Hall/CRC Press, Boca Raton
Guhaniyogi R, Finley AO, Banerjee S, Gelfand AE (2011) Adaptive Gaussian predictive process models for large spatial datasets. Division of Biostatistics, University of Minnesota. Research report rr2011-013
Hartigan JA, Wong MA (1979) A K-means clustering algorithm. Appl Stat 28(1):100–108
Higdon D Space and space-time modeling using process convolutions. In: Anderson C, Barnett C, Chatwin PC, El-Shaarawi AH (eds) Quantitative methods for current environmental issues.. pp 37–56 Springer, Berlin, (2002)
Kamman EE, Wand MP (2003) Geoadditive models. J R Stat Soc C 52(1):1–18
Kaufman CG, Schervish MJ, Nychka DW (2008) Covariance tapering for likelihood-based estimation in large spatial data sets. J Am Stat Assoc 103(484):1545–1555
Meehl G, Covey C, Delworth T, Latif M, McAvaney B, Mitchell J, Stouffer R, Taylor K (2007) The WCRP CMIP3 multimodel dataset: a new era in climate change research. Bull Am Meteorol Soc 88(9):1383–1394
Pace RK, LeSage JP (2009) A sampling approach to estimating the log determinant used in spatial likelihood problems. J Geograph Syst 11(3):209–225
Rougier J, Goldstein M, House L (2010) Assessing climate uncertainty using evaluations of several different climate simulators. Department of Mathematics, University of Bristol. Technical report
Salazar E, Sansó B, Finley AO, Hammerling D, Steinsland I, Wang X, Delamater P (2011) Comparing and blending regional climate model predictions for the American Southwest. Department of Applied Mathematics and Statistics, UC Santa Cruz. Technical report UCSC-SOE-11-14
Smith R, Tebaldi C, Nychka D, Mearns L (2009) Bayesian modeling of uncertainty in ensembles of climate models. J Am Stat Assoc 104(485):97–116
Stein ML, Chi Z, Welty LJ (2004) Approximating likelihoods for large spatial datasets. J R Stat Soc Ser B 66(2):275–296
Stein ML (2007) Spatial variation of total column ozone on a global scale. Ann Appl Stat 1(1):191–210
Stein ML (2008) A modeling approach for large spatial datasets. J Korean Stat Soc 37(1):3–10
Stroud JR, Muller P, Sansó B (2001) Dynamic models for spatiotemporal data. J R Stat Soc Ser B 63(4):673–689
Tebaldi C, Sansó B (2008) Joint projections of temperature and precipitation change from multiple climate models: a hierarchical Bayes approach. J R Stat Soc A 172(1):83–106
Tonellato S (1997) Bayesian dynamic linear models for spatial time series. Tech report, Rapporto di riceria 5/1997, Dipartimento di Statistica - Universita CaFoscari di Venezia, Venice Italy
Vecchia AV (1988) Estimation and model identification for continuous spatial processes. J R Stat Soc Ser B 50(2):297–312
West M, Harrison P (1997) Bayesian forecasting and dynamic models, 2nd edn. Springer, New York