Published in final edited form as:
Environmetrics. 2011 December ; 22(8): 997–1007. doi:10.1002/env.1131.

Adaptive Gaussian Predictive Process Models for Large Spatial Datasets

Rajarshi Guhaniyogi, **Andrew O. Finley**, **Sudipto Banerjee**, and **Alan E. Gelfand**[1]
Department of Forestry, Michigan State University, East Lansing, Michigan

Abstract

Large point referenced datasets occur frequently in the environmental and natural sciences. Use of Bayesian hierarchical spatial models for analyzing these datasets is undermined by onerous computational burdens associated with parameter estimation. Low-rank spatial process models attempt to resolve this problem by projecting spatial effects to a lower-dimensional subspace. This subspace is determined by a judicious choice of "knots" or locations that are *fixed* a priori. One such representation yields a class of *predictive process models* (e.g., Banerjee et al., 2008) for spatial and spatial-temporal data. Our contribution here expands upon predictive process models with *fixed* knots to models that accommodate stochastic modeling of the knots. We view the knots as emerging from a point pattern and investigate how such adaptive specifications can yield more flexible hierarchical frameworks that lead to automated knot selection and substantial computational benefits.

Keywords

Bayesian hierarchical models; Gaussian process; Intensity surfaces; Low-rank models; Markov chain Monte Carlo; Predictive process

1 Introduction

Recent developments in Geographical Information Systems (GIS) and Global Positioning Systems (GPS) enable accurate geocoding of locations where scientific data are collected. Today, large spatiotemporal datasets abound in many fields and have generated considerable interest in statistical models for location-referenced spatial data (Cressie, 1993; Banerjee, Carlin and Gelfand 2004; and Schabenberger and Gotway 2004). Estimating hierarchical spatial models involves matrix factorizations of the order of n^3 in the number of locations n – hence the infeasibility or "big n" problem for large datasets.

Low-rank or reduced-rank spatial models have become extremely popular for analyzing large spatial datasets (Stein, 2008; Cressie and Johannesson, 2008; Banerjee et al., 2008). These express the spatial effects over $\mathcal{S}$ in terms of its realizations over the smaller set of "knots", say $\mathcal{S}=\{s_1^*,\ldots,s_n^*\}$, where n^* is *fixed* to be much smaller than the number of observed sites. A key issue in low-rank methods is the choice of knots, which is usually dictated by computational cost and sensitivity to choice. In practice, we often investigate

Correspondence author: Andrew O. Finley telephone: (517) 432-7219 finleya@msu.edu .
[1]Rajarshi Guhaniyogi is a Ph.D. Candidate, School of Public Health at the University of Minnesota, Minneapolis, Minnesota, U.S.A., (guhan003@umn.edu); Andrew O. Finley is an Assistant Professor in the Departments of Forestry and Geography at the Michigan State University, East Lansing, Michigan, U.S.A., (finleya@msu.edu); Sudipto Banerjee is an Associate Professor of Biostatistics, School of Public Health at the University of Minnesota, Minneapolis, Minnesota, U.S.A., (baner009@umn.edu); Alan E. Gelfand is a Professor of Statistics and Decision Sciences, Duke University, Durham, North Carolina, U.S.A., (alan@stat.duke.edu)

sensitivity of inference to different choices of n^*, which entails separately estimating a number of low-rank models. Typically, for each n^* we use some space-covering design (e.g., Royle and Nychka, 1998) to fix the knots.

Our contribution here expands upon existing hierarchical low-rank models, as explored in the aforementioned references, to allow the knots to vary. While our formulation applies to any low rank likelihood, we specifically work with a flexible class called *predictive process* models. See Banerjee et al., (2008), Finley et al. (2009), and Eidsvik et al. (2010) for several methodological explorations and applications, but all with the knots fixed a priori. We assign a prior probability model for the knots to build a more automated low-rank hierarchical model and explore what benefits, if any, such stochastic modeling will fetch in terms of model performance and prediction.

The format of the manuscript is as follows. Section 2 provides a brief review of predictive process models. Section 3 comprises three subsections: Subsection 3.1 discusses stochastic modeling of the knots in our hierarchical setup, Subsection 3.2 discusses computational and implementation details, while Subsection 3.3 discusses spatial prediction and interpolation. Section 4 illustrates our adaptive models using a synthetic and forest inventory dataset. Finally, Section 5 concludes the manuscript with an eye towards future work.

2 Gaussian predictive process models – A brief review

Detailed descriptions of hierarchical Gaussian predictive process models are given in Banerjee et al. (2008), Finley et al. (2009), and Eidsvik et al. (2010). Here, we o er a brief review and introduce new notation that makes explicit the dependence on the knots. We envision an outcome variable $y(s)$ along with a $p\times 1$ vector of spatially referenced predictors $x(s)$, which are associated through a spatial regression model $\mathrm{E}[y(s) \mid x(s), \beta, w(s)] = x(s)'\beta + w(s)$. This includes a spatial process over the study region $\mathcal{D}$, defined by the set $w_{\mathcal{D}} = \{w(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ that can be viewed as a randomly realized surface (or a random function) over the region. A more succinct notation denotes the process as $w(s)$.

Customarily, $w(s)$ is specified as zero-centered Gaussian Process with a parametric covariance function $C(s_i, s_j; \theta) = \mathrm{cov}\{w(s_i), w(s_j)\}$ so that $[w(\mathcal{S})|\theta] = N(0, C(\mathcal{S};\theta))$, where $w(\mathcal{S}) = \{w(s_i) : s_i \in \mathcal{S}\}$ is a (partial) realization of $w(\cdot)$ over $\mathcal{S}$ and $C(\mathcal{S};\theta)$ is the $n \times n$ matrix whose (i, j)-th element is given by $C(s_i, s_j; \theta)$. Often we specify $C(s_i, s_j; \theta) = \sigma^2\rho(s_i, s_j; \boldsymbol{\phi})$, where $\theta = \{\sigma^2, \boldsymbol{\phi}\}$, σ^2 is a spatial variance component and $\rho(\cdot; \boldsymbol{\phi})$ denotes a spatial correlation function.

We avoid dealing with the $n \times n$ dense matrix $C(\mathcal{S};\theta)$, by projecting the process $w(s)$ at a generic location s onto a subspace spanned by its realization over $\mathcal{S}^*$. Under certain optimality criteria (see, e.g., Banerjee et al. 2008), we arrive at the *predictive process* $\tilde{w}(s) = \boldsymbol{c}(s, \mathcal{S}^*;\theta)' C(\mathcal{S}^*;\theta)^{-1} w(\mathcal{S}^*)$, where $c(s, \mathcal{S}^*;\theta)'$ is the $1 \times n^*$ vector with $c(s, s_j^*;\theta)$ as the j-th element. Finley et al. (2009) showed that the predictive process is smoother than the parent process, which detracts from its predictive performance. A remedy is to use the "bias-adjusted" predictive process $\tilde{w}_\epsilon(s) = \tilde{w}(s) + \tilde{\epsilon}(s)$, where $\tilde{\epsilon}(s) \overset{iid}{\sim} N\left(0, C(s, s;\theta) - c(s, \mathcal{S}^*;\theta)' C(\mathcal{S}^*;\theta)^{-1} c(s, \mathcal{S}^*)\right)$ and $\tilde{\epsilon}(s)$ is independent of $\tilde{w}(s)$.

Let $y = (y(s_1), y(s_2), \ldots, y(s_n))'$ be a vector of Gaussian outcomes. A hierarchical bias-adjusted predictive process model, conditional upon a fixed $\mathcal{S}^*$, can be written by marginalizing out the spatial effects $w(\mathcal{S}^*)$ to obtain the posterior distribution

$$\left[\beta,\theta,\tau^2|\boldsymbol{y},\mathcal{S}^*\right]\propto[\theta]\times IG\left(\tau^2|a_\tau,b_\tau\right)\times N\left(\beta|\mu_\beta,\Sigma_\beta\right)\times N\left(\mathbf{y}|\boldsymbol{X}\beta,\Sigma_{\boldsymbol{y}}\left(\mathcal{S}^*,\theta,\tau^2\right)\right). \quad (1)$$

Here X is $n \times p$ with $x(s_i)$' as the i-th row, $\Sigma_{\mathbf{y}}\left(\mathcal{S}^*,\theta,\tau^2\right)=C(\mathcal{S}^*;\theta)'C(\mathcal{S}^*;\theta)^{-1}C(\mathcal{S}^*;\theta)+D_{\tilde{\epsilon}+\epsilon}$ and $D_{\tilde{\epsilon}+\epsilon}=D_{\tilde{\epsilon}}+\tau^2\boldsymbol{I}_n$ with a diagonal matrix whose i-th diagonal element is the variance of $\tilde{\epsilon}(s_i)$. Letting $\boldsymbol{J}=\left[C(\mathcal{S}^*;\theta)+C(\mathcal{S}^*;\theta)D_{\tilde{\epsilon}+\epsilon}^{-1}C(\mathcal{S}^*;\theta)'\right]$, the Sherman-Woodbury-Morrison formulas (Henderson and Searle, 1981) yield $\Sigma_{\boldsymbol{y}}\left(\mathcal{S}^*,\theta,\tau^2\right)^{-1}$ as $D_{\tilde{\epsilon}+\epsilon}^{-1}-D_{\tilde{\epsilon}+\epsilon}^{-1}C(\mathcal{S}^*;\theta)'J^{-1}C(\mathcal{S}^*;\theta)D_{\tilde{\epsilon}+\epsilon}^{-1}$ and $|\Sigma_{\boldsymbol{y}}\left(\mathcal{S}^*,\theta,\tau^2\right)|$ as $\frac{|D_{\tilde{\epsilon}+\epsilon}|}{|C(\mathcal{S}^*;\theta)|}\times|J|$. These expressions involve inverses and determinants that are either diagonal or $n^* \times n^*$.

3 Inference about the Knots

Here we present two simple one-dimensional examples to illustrate why estimating the knot locations can be beneficial. The data used for the first example is shown in Figure 1(a) and (b). Here, the ○ symbols represent observed values of y, which were fixed at zero and the + symbols depict two knots located at the extremes of the x axis. Given these data, two models were used to predict the value of y for 100 new x values between 0 and 1. First, we use model (1) with only an intercept in the regression and the two fixed knots. Second, we let the knots vary by assigning a simple uniform prior $U(0, 1)$ for the position of each knot on the x axis. Posterior inference for each model was based on 5, 000 post burn-in Markov chain Monte Carlo (MCMC) samples. The median of each of the 100 posterior predictive distributions produced using the fixed and adaptive knot models are indicated with the • symbol in Figure 1(a) and (b), respectively. Here too, the associated posterior predictive 95% credible intervals are depicted with gray bands. The bottom density plot Figure 1(c) illustrates where the adaptive knots sampled. This density plot shows the adaptive knots move from the starting locations to locations that approximately divide the domain into equal portions. This is an intuitive and reassuring result, given that observations are equally distributed across the domain. Further, by comparing Figure 1(a) and (b) we can see the adaptive knot locations tend to produce narrower posterior predictive 95% credible intervals across the domain. For instance, the sum of the 100 95% credible interval ranges is 81 for the fixed knot model and 77 for the adaptive knot model. This trend is seen for other knot intensities. For example, the one and three knot intensities result in a decrease from 93 to 84 and 74 to 73 for fixed and adaptive knot models, respectively.

The second example follows the same setup, however, y is now drawn from a normal distribution with a varying frequency sine function mean and variance of 0.01. These data are show in Figure 2(a) and (b). Again, 100 new x values between 0 and 1 were withheld and used to assess the models' predictive performance. The seven knot locations, + symbols, were considered for the candidate models. These seven knots were again allowed to vary across the domain for the adaptive model.

The median of each of the 100 posterior predictive distributions produced using the fixed and adaptive knot models are indicated with the • symbol in Figure 2(a) and (b), respectively. Here too, the associated posterior predictive 95% credible intervals are depicted with gray bands. Prediction using the fixed knot model is based only on information at the knot locations and, as a result, produces a poor approximation of y, as reflected by the results in Figure 2(a). In contrast, Figure 2(b) shows that by allowing the knot locations to move along the x axis, and learn from the data, predictions from the

adaptive knot model more accurately capture the distribution of *y*. The bottom density plot Figure 2(c) illustrates where the adaptive knots sampled. This plot shows that knots tend to sample at values of *x* that correspond to the stationary points on the sine curve, resulting in a greatly improved approximation of the original data.

The adaptive knot model involves an averaging over the distribution of the knots, which often, at least in practice, leads to reduced predictive variability. This can be seen when comparing the coverage and width of the 95% credible interval bands in Figures 1 and 2. It is worth pointing out that some form of an averaged predictive variance over the spatial domain is often used as an objective function to resolve optimal spatial design (e.g., Diggle and Lophaven, 2006; Zhu and Stein, 2005).

3.1 Modeling the knots

Fixing the number of knots, n^*, based upon available computing resources, we allow knot locations to vary, conditional upon $\eta_D = \{\eta(s) : s \in D\}$, where $\eta(s) = \exp(\lambda(s))$ is an intensity function, according to the density

$$[\mathcal{S}^*|\eta_D, n^*] = \prod_{i=1}^{n^*} \frac{\eta\left(s_i^*\right)}{\int_{\mathcal{D}} \eta(s)\, ds} = \left(\int_{\mathcal{D}} \eta(s)\, ds\right)^{-n^*} \times \prod_{i=1}^{n^*} \eta\left(s_i^*\right). \tag{2}$$

We extend (1), allowing the knot locations to vary over $\mathcal{D}$, by

$$\begin{aligned}\left[\beta, \tau^2, \theta_1, \quad \mathcal{S}^*, \theta_2|\mathbf{y}, \mathcal{S}, n^*\right] \propto [\theta_1] \times IG\left(\tau^2|a_\tau, b_\tau\right) \times N\left(\beta|\mu_\beta, \Sigma_\beta\right) \\ \times [\eta_{\mathcal{D}}] \times [\mathcal{S}^*|\eta_{\mathcal{D}}] \times N\left(\mathbf{y}|\boldsymbol{X}\beta, \Sigma_{\boldsymbol{y}}\left(\mathcal{S}^*; \theta_1, \tau^2\right)\right),\end{aligned} \tag{3}$$

where θ_1 now represents the process parameters in the data likelihood. Two practical approaches for modeling $[\eta_D] \times [S^* \mid \eta_D]$ are outlined below.

3.1.1 Modeling $\eta(s)$ - a parametric model—Parametric forms can be prescribed for $\eta(s)$, such as basis representations or tiled surfaces (see, e.g., Diggle, 2003). Here, we employ a random equally weighted bivariate normal mixture, and then add priors on the parameters in the normal kernel, say θ_2. More specifically, let $\theta_2 = \{u_1, u_2, \ldots, u_m, \sum_\eta\}$, where u_j's are m points in $\mathcal{D}$, $\sum_\eta$ is a common 2×2 variance covariance matrix. The intensity is $\log\{\eta(s;\theta_2)\} = \frac{1}{m}\sum_{j=1}^{m} N_{2\mathcal{D}}\left(s|u_j, \Sigma_\eta\right)$, where $N_{2D}(\cdot \mid u_j, \sum_\eta)$ denotes a bivariate normal density, truncated to $\mathcal{D}$, with mean u_j and variance-covariance matrix $\sum_\eta$. This parametric kernel specification replaces $[\eta_D] \times [\mathcal{S}^* \mid \eta_D]$ with $[\theta_2] \times [\mathcal{S}^* \mid \theta_2]$ in (3) (suppressing the implicit conditioning on n^*).

Prior specifications for θ_2 typically comprise a uniform support over $\mathcal{D}$ for each of the u_j's and an inverse Wishart $IW(r_\eta, \Omega_\eta)$ (e.g., Gelman et al. 2004) for $\sum_\eta$. Alternatively, we1 could further parametrize $\Sigma_\eta = \sigma_\eta^2 \begin{pmatrix} 1 & \rho_\eta \\ \rho_n & 1 \end{pmatrix}$ and assign appropriate priors to σ_η^2 and ρ_η.

3.1.2 Modeling $\eta(s)$ - a log-Gaussian model—Rather than the parametric choice above, we can use a log-Gaussian process $\eta(s) = \exp\{w_2(s)\}$ where $w_2(s)$ is a Gaussian process with zero mean, unit variance and correlation function $\rho_2(\cdot; \phi_2)$. There remains an analytically inaccessible integral of $w_2(s)$. Matters are assuaged by a lower-dimensional

representation for $w_2(s)$. One option, naturally, is the predictive process itself and we replace $w_2(s)$ by $\tilde{w}_2(s)$.

More specifically, let $\{u_1, u_2, \ldots, u_m\}$ be a set of *fixed* knots and let $\tilde{w}_2(s) = \mathrm{E}\left[w_2(s) | w_2^*\right]$ be the corresponding predictive process, where $w_2^* = \left(w_2(u_1), w_2(u_2), \ldots, w_2(u_m)\right)'$. The corresponding hierarchical model is still embodied by (3) with $\theta_2 = \left\{\phi_2, w_2^*\right\}$ and $[\theta_2] = [\phi_2] \times N_m\left(w_2^* | 0, R_2(\phi_2)\right)$, where $\boldsymbol{R}_2(\boldsymbol{\phi}_2)$ is the $m \times m$ correlation matrix with $\rho_2(u_i, u_j; \boldsymbol{\phi}_2)$ as the (i, j)-th element. Our experiments reveal that a modest value of m, usually between 10 and 30, allows adequate exploration of most domains by the knots. For a given size of n^*, increasing m beyond ~10 did not substantially alter the final inference.

3.2 Implementation details

The parameters in (3) are updated using a combination of Gibbs and Metropolis steps. We first update β from $N(\mu_{\beta|\cdot}, \Sigma_{\beta|\cdot})$, with covariance matrix $\Sigma_{\beta|\cdot} = \left(\boldsymbol{X}'\Sigma_y\left(S^*; \theta_1, \tau^2\right)^{-1} X + \Sigma_\beta^{-1}\right)^{-1}$, and mean $\mu_{\beta|\cdot} = \left(\boldsymbol{X}'\Sigma_y\left(S^*; \theta_1, \tau^2\right)^{-1} X + \Sigma_\beta^{-1}\right)^{-1}$ The inverse of $\Sigma_y(S^*; \theta_1, \tau^2)$ is obtained from the Sherman-Woodbury-Morrison formulas.

We update $\{\theta_1, \tau^2\}$, $\{S^*\}$ and $\{\theta_2\}$ in separate blocks. The step for θ_2 requires some clarification. This involves evaluating $[\theta_2] \times [S^* \mid \theta_2]$, which involves the integral $\int_{\mathcal{D}} \eta(s; \theta_2)\, ds$ in (2). We use a grid-based integration scheme. Letting $\upsilon_1, \upsilon_2, \ldots, \upsilon_M \in \mathcal{D}$ be the grid of points covering $\mathcal{D}$ and each cell area equal to Δ, we approximate $\int_{\mathcal{D}} \eta(s; \theta_2)\, ds \approx \Delta \sum_{j=1}^{M} \eta\left(v_j; \theta_2\right)$. The full conditional distribution for $\theta_2 = \left[\left\{u_j\right\}_{j=1}^{m}, \Sigma_\eta\right]$ depends upon how we model $\eta(s; \theta_2)$. For the bivariate parametric kernels, as in Section 3.1.1 it is proportional to

$$\prod_{j=1}^{m}\left[\boldsymbol{u}_j | \mathcal{D}\right] \times IW\left(\Sigma_\eta | r_n, \Omega_\eta\right) \times \left(\sum_{j=1}^{M} \eta\left(\mathbf{v}_j; \theta_2\right)\right)^{-n^*} \times \prod_{i=1}^{n^*} \eta\left(s_i^*; \theta_2\right). \tag{4}$$

A common choice for each $\left[u_j | \mathcal{D}\right]$ is a bivariate uniform density over $\mathcal{D}$.

When the log-Gaussian process, as described in Section 3.1.2, is used to model S^*, the only change in parameters arises in θ_2, which now comprises $\left\{\phi_2, w_2^*\right\}$. The full conditional distribution for θ_2 is proportional to

$$[\phi_2] \times N_m\left(\boldsymbol{w}_2^* | 0, \boldsymbol{R}_2(\phi_2)\right) \times \left(\int_{\mathcal{D}} \eta(\boldsymbol{s}; \theta_2)\, d\boldsymbol{s}\right)^{-n^*} \times \prod_{i=1}^{n^*} \eta\left(s_i^*; \theta_2\right). \tag{4′}$$

Evaluating (4′) entails approximating the integral of the intensity surface in each iteration. We conveniently take the u_i's in Section 3.1.2 over a grid and use the current estimate for θ_2 to approximate $\int_{\mathcal{D}} \eta(s; \theta_2)\, ds \approx \Delta \sum_{j=1}^{m} \eta\left(u_j; \theta_2\right)$. Details regarding the choice of priors and updating steps are provided in Section 4 in the context of specific examples.

3.3 Spatial prediction, interpolation and model assessment

For predicting $Y(s_0)$ at any location s_0 in the domain, we sample from the posterior predictive distribution, $[Y(s_0) \mid y] = \int [Y(s_0) \mid y; \theta_1, \tau^2, S^*][\theta_1, \tau^2, S^* \mid y]$ using *composition* (e.g., Banerjee et al. 2004). For each $\{\theta_1^{(l)}, \tau^{2(l)}, \mathcal{S}^{*(l)}\}$ for $l = 1, 2,...,L$, obtained from the posterior distribution $[\theta_1; \tau^2; S^* \mid y]$, we draw $Y(s_0)^{(l)}$ from $\left[Y(s_0) \mid y, \theta_1^{(l)}, \tau^{2(l)} \mathcal{S}^{*(l)}\right]$. For inference on the spatial process, $\tilde{w}_{\tilde{\epsilon}}(s_0)$, we use posterior predictive samples from

$$[\tilde{w}_{\tilde{\epsilon}}(\boldsymbol{s}_0) \mid \boldsymbol{y}] = \int \left[\tilde{w}_{\tilde{\epsilon}}(\boldsymbol{s}_0) \mid w(\mathcal{S}^*), \theta_1, \tau^2, \mathcal{S}^*\right] \left[w(\mathcal{S}^*) \mid \boldsymbol{y}, \theta_1, \tau^2, \mathcal{S}^*\right] [\theta_1, \mathcal{S}^* \mid \boldsymbol{y}]\, d\theta_1 d\mathcal{S}^* dw(\mathcal{S}^*)\, d\tau^2.$$

We first sample $\{\theta_1^{(l)}, \tau^{2(l)}, \mathcal{S}^{*(l)}\}$, for $l = 1, 2,\ldots,L$, from the posterior distribution. Next, we sample $w(S^*)^{(l)}$ from $\left[w(\mathcal{S}^*) \mid y, \theta_1^{(l)}, \tau^{2(l)}, \mathcal{S}^{*(l)}\right]$ which, in fact, is a normal distribution, and finally, we sample $\tilde{w}_{\tilde{\epsilon}}(s_0)^{(l)}$ from $[w(s_0) \mid w(S^*)^{(l)}, \theta^{(l)}, S^{*(l)}]$, again, a normal distribution.

We assess model performance using *independent* replicates for each observed outcome: for each $s_i \in S$, we draw $Y_{rep}(s_i)^{(l)}$ from $N\left(x(s_i)'\beta^{(l)} + \tilde{w}_\epsilon(s_i)^{(l)}, \tau^{2(l)}\right)$, one for one for the posterior samples. Letting $\mu_{rep,i}$ and $\sigma^2_{\text{rep},i}$ be the posterior predictive mean and variance for each $Y_{rep}(s_i)$, we compute $G = \sum_{i=1}^{n} \left(y(s_i) - \mu_{\text{rep},i}\right)^2$ and $P = \sum_{i=1}^{n} \sigma^2_{\text{rep},i}$. We use $D = G + P$ (e.g., Gelfand and Ghosh, 1998) as a model selection criteria, [summationtext] with lower values of D indicating better models. Further, for each analysis we used a holdout set to assess each models' predictive performance by computing the mean squared prediction error (MSPE), $\frac{1}{q}\sum_{i=1}^{q} \left(y(s_i) - \tilde{Y}(s_i)\right)^2$, where $\tilde{Y}(s_i)$ is the predicted outcome at the i-th holdout location and q is the number of locations in the holdout set.

4 Illustrations

We use both a synthetic and forest inventory dataset to assess model performance with regard to learning about process parameters and predicting at new locations. Posterior inferencewas based on three chains of 25, 000 iterations (the first 5,000 iterations were discarded as burn-in). The samplers were coded in C++ and Fortran and leveraged Intel's Math Kernel Library threaded BLAS and LAPACK routines for matrix computations. All analyses were conducted on a Linux workstation using two Intel Nehalem quad-Xeon processors.

4.1 Synthetic data analysis

The synthetic dataset comprises $n = 5,500$ observations within a unit square domain with outcome values generated from $N(\beta_0\mathbf{1}, \sigma^2\boldsymbol{R}(\boldsymbol{\phi}) + \tau^2\boldsymbol{I})$ with $\boldsymbol{R}(\phi)$ an $n \times n$ correlation matrix whose (i, j)-th element is $e^{-\phi d_{ij}}$, $d_{ij} = \|s_i - s_j\|$ and parameters given in the first column of Table 1. Figure 3(a) illustrates the spatial random e ect surface interpolated over the $w(s)$'s. To facilitate model comparison using predictive performance, 500 observations were withheld to serve as a holdout set. Eleven gridded knot intensities ($n^* = (5^2, 6^2, \ldots, 15^2)$ were considered for both the non-adaptive (i.e., fixed knot) and adaptive bias-adjusted predictive process models. For all models, the intercept parameter β_0 was given a *flat* prior and the variance parameters τ^2 and σ^2 each received inverse-Gamma $IG(2, 1)$ priors. Further, assuming an exponential spatial correlation function the prior for the spatial decay parameter ϕ was a Uniform $U(3, 300)$, which corresponds to support between 0.01 and 1.0 in map

distance units. This is a broad range of support considering the maximum distance between any two observations is 1.4. The adaptive knot models follow the log-Gaussian parameterization of $\eta(s)$, detailed in Subsection 3.1.2, with a broad prior support of $U(3, 300)$ on ϕ_2.

Results for the 25, 36, 196, and 225 knot models are detailed in Table 1. Here, both the non-adaptive and adaptive models produce similar estimates of β_0 across the range of knot intensities. When n^* is small, the sparse grid of knots provides an over smoothed representation of the latent spatial surface and, as a result, the non-adaptive model is not able to accurately estimate the spatial random effect parameters σ^2 and ϕ. In contrast, even at a 25 knot intensity, the adaptive model provides better estimates of these parameters. Note, however, that the nugget, τ^2, is apparently underestimated for the adaptive model. In fact, the bias-adjustment incorporated here may tend to slightly over-fit. It is not surprising, therefore, that the adaptive knots further overcompensate for the bias, which, after all, is characterized only for the fixed-knot setting. The posterior predictive variances for the response at each site is not, in our experience, affected substantially by this bias.

Considering the non-adaptive and adaptive models' D and MSPE across knot intensities in Table 1, it is clear that knot location influences model fit and subsequent prediction. For example, with just 25 knots, the adaptive model produced D=45880. This level of fit was not achieved until the ~81 knot intensity for the non-adaptive model. In addition to a consistently better model fit (i.e., lower D), the adaptive model offers considerably lower MSPE across all knot intensities. For instance, the 9.24 MSPE of the 225 knot non-adaptive model is considerably larger than the 9.05 MSPE achieved by the 36 knot adaptive model.

Even with the reduced dimensionality afforded by the predictive process, fitting these models is time consuming. The last row in each section of Table 1 gives the run time for 25,000 MCMC iterations on a single non-hyperthreaded processor. The extra complexity of the adaptive predictive process sampler approximately doubles the run time across all knot intensities. Importantly, however, the adaptive model can produce an MSPE of 9.05 in 12 hours versus the non-adaptive model's substantially larger 9.24 MSPE of the 225 knot model, which requires a 28.5 hour run time. Further, the 225 knot non-adaptive model's fit, of D=37007, is achieved by the 81 knot adaptive model, which had a run time of 23.0 hours.

The plots in Figure 3 can help us understand the adaptive model's advantage over the fixed knot model. Figure 3(b) is a trace plot of knot movement for the adaptive 25 knot model. Here, the • symbols mark the states of one MCMC chain over the domain. The density surface associated with Figure 3(b) is illustrated in Figure 3(c), where higher values (darker shades) indicate the regions where the knots were sampled more intensely. By comparing the *true* spatial random effect surface Figure 3(a) with Figure 3(c) it is apparent that the knots tend to move to regions of extreme $w(s)$ values (as seen in the one-dimensional example offered in Section 1). This is a trend repeated across the adaptive predictive process models. Figures 3(d) and (e) were generated by interpolating over the median of each location's spatial random effect posterior distribution calculated using the non-adaptive and adaptive 25 knot predictive process models, respectively. Comparing these surfaces with Figure 3(a) shows that the adaptive knots provide a more detailed representation of the spatial random effect surface, hence improved model fit and predictive ability.

Given the trade-off between the non-adaptive and adaptive knot models' run time and predictive performance, we considered a hybrid scheme that used the knot locations of the 1000-th MCMC iteration of an adaptive model. Here, the choice of the 1000-th iteration was arbitrary; however, the idea was to allow enough iterations for the knots to move about the domain, while keeping the run time to a minimum. As summarized in the last column in the

second row of Table 1, this hybrid model seems to enjoy the improved fit and lower MSPE of the adaptive model and the shorter run time advantage of the non-adaptive model.

4.2 Forest biomass data analysis

Spatial prediction of forest biomass is critical to many important contemporary global-, regional-, and local-scale decisions, including assessments of current carbon stock and flux, bio-feedstock for emerging bio-economies, and impact of deforestation. In the United States, the Forest Inventory and Analysis (FIA) program of the USDA Forest Service collects the data needed to support these assessments.

The program has established field plot centers in permanent locations using a sampling design that produces an equal probability sample (Bechtold and Patterson, 2005). Locations of the 7.32 m radius forested plots are determined using GPS receivers. The state of Michigan, in which the study area is located, has a sampling intensity of approximately one plot per 800 ha. On these plots, field crews recorded stem measurements for all trees with diameter at breast height (dbh; 1.37 m above the forest floor) of 12.7 cm or greater. Given these data, established allometric equations were used to estimate each plot's forest biomass per ha. Here, we model the log metric tons of forest biomass per ha. A July, 2003 mosaic of Landsat TM imagery, was used to calculate tasseled cap components of brightness (TC1), greenness (TC2), and wetness (TC3) to serve as predictor variables (Huang et al., 2002). Figure 4(a) illustrates the georeferenced forest inventory data consisting of 6,538 forested FIA plots measured between 1999 and 2006 across the lower peninsula of Michigan.

Candidate models include a simple non-spatial regression and the non-adaptive and adaptive bias-adjusted predictive process models. Similar to the synthetic data analysis, we considered a range of knot intensities. Knot locations were chosen by applying a *k-means* clustering algorithm (Hartigan and Wong, 1979) to the observed locations. For the adaptive models these knot locations served as starting values. As in Section 4.1, we considered an additional non-adaptive candidate model, that used the knot locations of the 1000-th MCMC iteration of an adaptive model.

Based on results from an initial variogram analysis of the non-spatial model's residuals, the priors for τ^2 and σ^2 for the non-adaptive and adaptive predictive process models followed *IG*(2 0.5). Assuming an Exponential spatial correlation function the prior for the spatial decay parameter ϕ followed a *U*(0.006 3), which corresponds to support from 1–500 km. Again, this is a broad range of support, given the maximum distance between any two plots is 460 km. For all models the regression coeffcients each received a *flat* prior. The prior on the adaptive knot locations, $\mathcal{S}^*$, was defined by a rectangular domain that covered the extent of the irregularly shaped study area.

Here, we opted to use the parametric parameterization of $\eta(\mathbf{s})$ detailed in Section 3.1.1. Priors for the parameters comprising the bivariate normal mixture covariance matrix $\sum_\eta$ were an *IG*(2, 1) for σ^2_η and *U*(−1, 1) for ρ_η. The mixture was evaluated over a grid of m=25 locations. We experimented with a range of m, 25–100, and found that it had negligible influence on parameter estimates and subsequent prediction.

Candidate models were assessed based on their fit to observed data, predictive performance at new locations, and run time. To assess predictive performance, 653 observations (i.e., 10%) were selected randomly to serve as a holdout set. The remaining 5,885 observations were used to fit the candidate models.

Figure 4(b) is an interpolated surface of the non-spatial model residuals. We would expect the fitted spatial random effects of the candidate models to look somewhat similar to this

residual surface. Figure 4(c) provides the density plot of the adaptive knot locations over25,000 MCMC iterations for the 50 knot model. Here, darker colors correspond to regions where the knots sampled more intensely and the • symbols indicate the starting location of the 50 knots. As in Section 4.1, the knots moved from the starting locations to sample at locations where the absolute value of the residual surface is large. This figure also shows the knots generally sample at locations close to the observed data (i.e., they did not sample much beyond the state's bounding polygon). Figure 4(d) and (e) provide interpolated surfaces of the median of spatial random effects posterior distribution for the adaptive and non-adaptive models, respectively. Here, we see the adaptive model produces spatial random effects that more closely approximate Figure 4(b) and hence provide improved model fit and prediction over the fixed knot model as detailed in Table 2.

Table 2 shows parameter estimates for candidate models with different knot intensities. Both the non-adaptive and adaptive predictive process models produce comparable estimates of the regression coe cients – several of which explain a significant amount of variability in log biomass. We again see a discrepancy between non-adaptive and adaptive models' estimates of the variance components. Specifically, the non-adaptive model seems to estimate a larger nugget, τ^2, and a smaller partial sill, σ^2, whereas the opposite trend appears in the adaptive model. This was also observed in Section 4.1 and other exploratory analysis we conducted. Results suggest that ~100 fixed knots are needed to produce MSPE and model fit, D, comparable to that of the 50 knot adaptive model. However, this advantage is lessened by the near equal run times of the two models (as noted in the last row of Table 2). The last column in this table presents the results for the non-adaptive model that used the knot locations of the 1000-th MCMC iteration of the 50 knot adaptive model. Similar to the synthetic analysis, this hybrid model offers the improved fit and predictive ability of an adaptive knot model with a run time comparable to that of the fixed knot model.

5 Discussion and future work

The current manuscript integrates modeling of knots in low rank predictive process models within a hierarchical framework, thereby circumventing issues underlying the choice of "knots". Indeed, we were able to obtain essentially indistinguishable inference with fewer number of stochastic knots than with fixed knots. Also, our approach applies seamlessly to other low-rank models that use kernel convolutions or other nonstationary covariance structures (e.g., Higdon, 2002; Cressie and Johannesson, 2008).

In the fixed knot setting, adding the bias-adjustment to the predictive process usually reduces smoothing and yields improved parameter estimates (Banerjee et al., 2010). However, the results of the analyses presented here show the bias-adjusted adaptive predictive process slightly underestimates τ^2. Remedies for avoiding such over-fitting could be achieved by tapering the bias-adjustment using a tapered covariance function (e.g., Furrer et al., 2006). Feasibility of alternative estimation strategies such as INLA (Rue et al., 2009; Eidsvik et al., 2010) can also be explored.

Any random probability measure for $[\mathcal{S}^*|\theta_2]$ will yield a valid hierarchical model in (3). If we eschew spatially informative priors for the knots, a fully non-parametric option using realizations from a Dirichlet Process will be viable. Algorithms to estimate such models have been outlined, among others, by Neal (1998). Stochastic modeling for random locations also arise when the outcome is "preferentially sampled" (Diggle et al., 2010), which seeks joint modeling for the process and the set of observed locations $\mathcal{S}$. Pati et al., (2011) recently proposed a hierarchical Bayesian geostatistical model for preferentially sampled data. An adaptive predictive process can be envisioned through distributions for $[\mathcal{S}|\theta_1]$ in (3) that would add an additional level of hierarchy. Finally, one can compare the

performance of the adaptive predictive model with different design-based strategies for fixed knot selection as outlined in recent work by Gelfand, Banerjee and Finley (2011). Such explorations will include random sampling strategies as well as further investigations into sensible "hybrid schemes" that combine adaptive strategies to arrive at fixed knot configurations. These constitute some ongoing projects.

References

Banerjee, S.; Carlin, BP.; Gelfand, AE. Hierarchical Modeling and Analysis for Spatial Data. Chapman and Hall/CRC Press; Boca Raton, FL: 2004.

Banerjee S, Gelfand AE, Finley AO, Sang H. Gaussian predictive process models for large spatial datasets. Journal of the Royal Statistical Society Series B. 2008; 70:825–848.

Banerjee S, Finley AO, Waldmann P, Ericsson T. Hierarchical spatial process models for multiple traits in large genetic trials. Journal of the American Statistical Association. 2010; 105:506–521. [PubMed: 20676229]

Bechtold, WA.; Patterson, PL., editors. The enhanced Forest Inventory and Analysis National Sample Design and Estimation Procedures. SRS-80. U.S. Department of Agriculture, Forest Service, Southern Research Station; Asheville, NC: 2005.

Cressie, NAC. Statistics for Spatial Data. 2nd edition. Wiley; New York: 1993.

Cressie NAC, Johannesson G. Spatial prediction for massive datasets. Journal of the Royal Statistical Society Series B. 2008; 70:209–226.

Diggle, PJ. Statistical Analysis of Spatial Point Patterns. Second edition. Arnold, London: 2003.

Diggle P, Lophaven S. Bayesian geostatistical design. Scandinavian Journal of Statistics. 2006; 33:53–64.

Diggle P, Menezes R, Su T. Geostatistical inference under preferential sampling. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2010; 59:191–232.

Eidsvik, J.; Finley, AO.; Banerjee, S.; Rue, H. Technical report. Norwegian University of Science and Technology; 2010. Approximate Bayesian inference for large spatial datasets using predictive process models.

Finley AO, Sang H, Banerjee S, Gelfand AE. Improving the performance of predictive process modeling for large datasets. Computational Statistics and Data Analysis. 2009; 53:2873–2884. [PubMed: 20016667]

Furrer R, Genton MG, Nychka D. Covariance tapering for interpolation of large spatial datasets. Journal of Computational and Graphical Statistics. 2006; 15:502–523.

Gelfand, AE.; Banerjee, S.; Finley, AO. Spatial design for knot selection in knot-based dimension reduction models. In: Mateu, J.; Muller, W., editors. Spatio-temporal Design: Advances in Efficient Data Acquisition. John Wiley; Chichester, UK: 2011.

Gelfand AE, Ghosh SK. Model Choice: a minimum posterior predictive loss approach. Biometrika. 1998; 85:1–11.

Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. Bayesian Data Analysis. 2nd edition. Chapman and Hall/CRC Press; Boca Raton, FL: 2004.

Hartigan JA, Wong MA. A K-means clustering algorithm. Applied Statistics. 1979; 28:100–108.

Henderson HV, Searle SR. On deriving the inverse of a sum of matrices. SIAM Review. 1981; 23:53–60.

Higdon, D. Space and Space-Time Modeling Using Process Convolutions. In: Anderson, C.; Barnett, V.; Chatwin, PC.; El-Shaarawi, AH., editors. Quantitative methods for current environmental issues. Springer-Verlag; 2002. p. 37-56.

Huang C, Wylie B, Homer C, Yang L, Zylstra G. Derivation of a tasseled cap transformation based on Landsat 7 at-satellite reflectance. International Journal of Remote Sensing. 2002; 8:1741–1748.

Neal, RM. Technical Report No. 9815. Department of Statistics, University of Toronto; 1998. Markov chain sampling methods for Dirichlet process mixture models.

Pati D, Reich BJ, Dunson DB. Bayesian geostatistical modeling with informative sampling locations. Biometrika. 2011; 98:35–48.

Royle JA, Nychka D. An algorithm for the construction of spatial coverage designs with implementation in SPLUS. Computers and Geosciences. 1998; 24:479–88.

Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). Journal of the Royal Statistical Society, Series B. 2009; 71:1–35.

Schabenberger, O.; Gotway, CA. Statistical Methods for Spatial Data Analysis. Chapman and Hall/CRC Press; Boca Raton, FL: 2004.

Stein ML. A modeling approach for large spatial datasets. Journal of the Korean Statistical Society. 2008; 37:3–10.

Zhu Z, Stein M. Spatial sampling design for parameter estimation of the covariance function. Journal of Statistical Planning and Inference. 2005; 134:583–603.

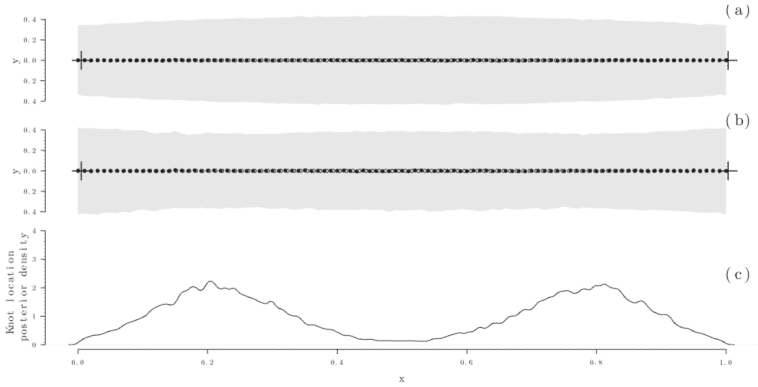

Figure 1.
Knot inference and prediction results for a fixed and adaptive knot model, (a) and (b) respectively. Models fit using observed data (○) and knot locations (+). Models' estimated posterior predictive median (•) and associated 95% credible interval (gray band) for 100 new locations. Adaptive knot model's posterior density of knot locations illustrated in (c).

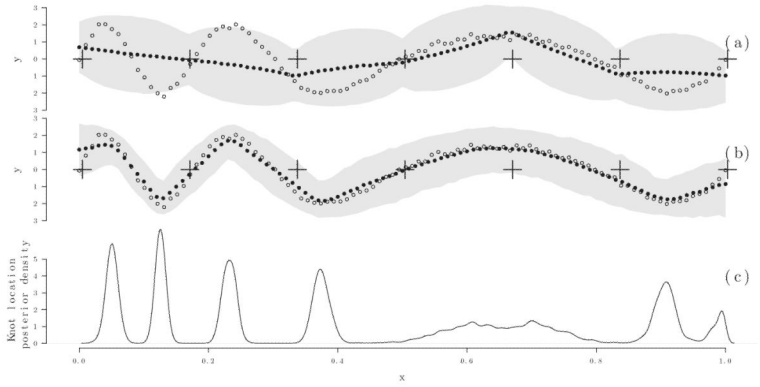

Figure 2.

Knot inference and prediction results for a fixed and adaptive knot model, (a) and (b) respectively. Models fit using observed data (○) drawn from a normal distribution with a varying frequency sine function mean and knot locations (+). Models' estimated posterior predictive median (•) and associated 95% credible interval (gray band) for 100 new locations. Adaptive knot model's posterior density of knot locations illustrated in (c).

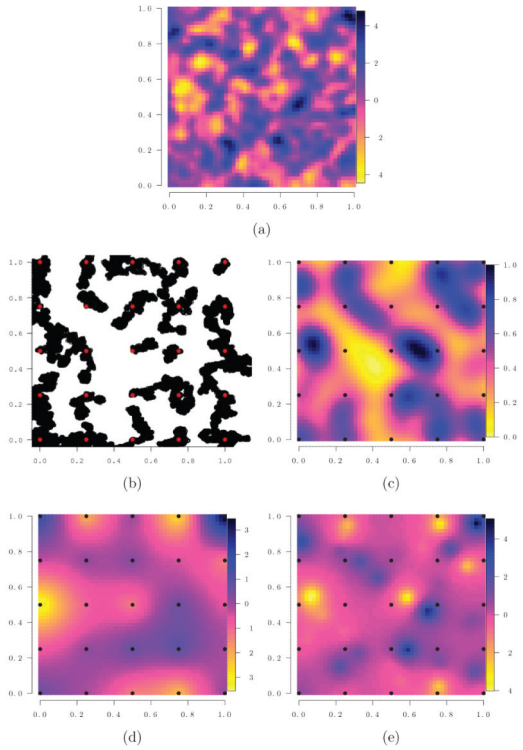

Figure 3.
Synthetic data and associated estimates for the 25 knot predictive process models: (a) synthetic spatial random effect surface generated using 5,000 observations; (b) 25,000 MCMC iteration trace plot of the adaptive knot locations; (c) density plot associated with the MCMC iteration in (b); (d) non-adaptive predicted process model estimated spatial random effects, and (e) adaptive predicted process model estimated spatial random effects.

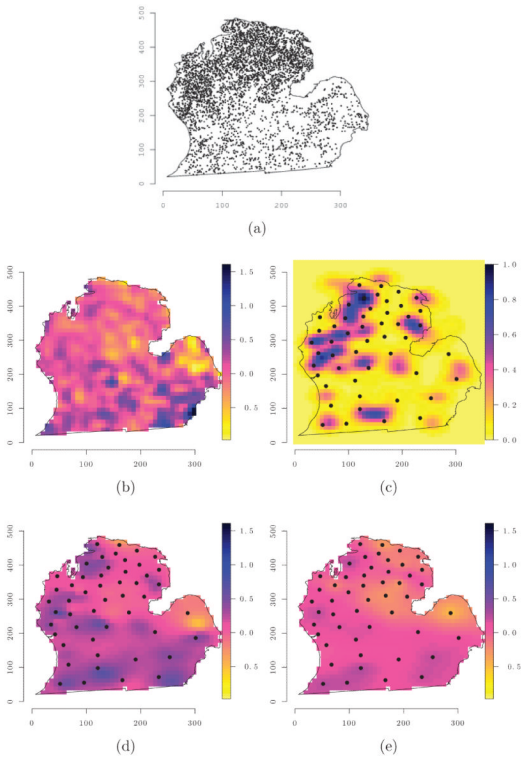

Figure 4.
Forest biomass dataset and associated estimates for the 50 knot predictive process models: (a) location of forest invetory plots; (b) interpolated surface of the non-spatial model residuals; (c) density plot of the adaptive knot locations over 25,000 MCMC iterations; (d) adaptive predicted process model estimated spatial random effects with knot starting locations, and; (e) non-adaptive predicted process model estimated spatial random effects with knot locations.

Table 1

Predictive process candidate models' parameter posterior credible intervals 50 (2.5 97.5), model fit criterion, and mean squared prediction error (MSPE) for the synthetic dataset. Run time is for a single chain of 25,000 iterations on a single non-hyperthreaded processor.

		Non-adaptive predictive process (i.e., fixed knots)			
	True	**25**	**36**	**196**	**225**
β_0	1	1.62 (0.41, 2.91)	1.10 (−0.02, 2.58)	1.22 (0.56, 1.90)	1.26 (0.64, 1.91)
σ^2	5	2.47 (1.62, 3.49)	2.91 (1.93, 4.33)	4.94 (4.04, 6.43)	5.22 (4.45, 6.41)
τ^2	1	4.18 (3.82, 4.49)	3.91 (3.43, 4.35)	1.76 (1.28, 2.16)	1.45 (1.03, 1.79)
ϕ	30	3.45 (3.02, 5.28)	4.26 (3.11, 7.32)	14.02 (9.76, 18.06)	15.11 (11.41, 18.75)
G	–	25123	24725	18762	17726
P	–	25545	25145	20068	19281
D	–	50669	49871	38831	37007
MSPE	–	10.37	10.22	10.11	9.24
Run time (hours)	–	5.0	6.0	24.5	28.5

		Adaptive predictive process				Non-adaptive with adaptive starting
	True	**25**	**36**	**196**	**225**	**36**
β_0	1	1.24 (1.15, 1.34)	1.29 (1.18, 1.40)	1.05 (0.83, 1.31)	1.11 (0.91, 1.52)	0.70 (0.51, 0.86)
σ^2	5	4.60 (4.37, 4.87)	4.40 (4.10, 4.70)	4.63 (4.32, 4.92)	4.69 (4.40, 5.07)	4.84 (4.50, 5.13)
τ^2	1	0.27 (0.13, 0.43)	0.30 (0.12, 0.50)	0.14 (0.08, 0.22)	0.13 (0.08, 0.26)	0.35 (0.21, 0.61)
ϕ	30	23.31 (21.13, 26.06)	22.30 (20.54, 24.51)	24.46 (22.16, 27.01)	24.08 (21.47, 26.93)	18.55 (16.65, 20.34)
ϕ_2	–	155.34 (23.11, 290.71)	160.39 (18.05, 290.93)	3.14 (3.00, 3.95)	3.20 (3.02, 3.90)	–
G	–	22722	21032	15395	14782	23025
P	–	23157	21825	17615	17186	23464
D	–	45880	42858	33010	31968	46490
MSPE	–	9.66	9.05	7.06	6.84	9.69
Run time (hours)	–	10.0	12.01	50.0	59.4	6.6

Table 2

Predictive process candidate models' parameter posterior credible intervals 50 (2.5 97.5), model fit criterion, and mean squared prediction error (MSPE) for the forest biomass dataset. Run time is for a single chain of 25,000 iterations on a single non-hyperthreaded processor.

	Non-spatial	Non-adaptive predictive process			Adaptive predictive process	Non-adaptive with adaptive starting
		50	100	200	50	50
β_0	10.98 (10.95, 11.01)	10.99 (10.71, 11.26)	11.00 (10.68, 11.30)	11.01 (10.97, 11.06)	10.81 (10.69, 10.94)	10.92 (10.82, 11.04)
β_{TC1}	0.07 (0.02, 0.12)	0.03 (−0.02, 0.09)	0.03 (−0.02, 0.09)	0.05 (0.00, 0.11)	0.06 (0.00, 0.11)	0.04 (−0.02, 0.09)
β_{TC2}	−0.03 (−0.09, 0.02)	−0.00 (−0.07, 0.06)	−0.01 (−0.07, 0.06)	−0.02 (−0.08, 0.04)	−0.02 (−0.09, 0.02)	−0.01 (−0.07, 0.05)
β_{TC3}	0.45 (0.41, 0.49)	0.43 (0.39, 0.48)	0.43 (0.39, 0.48)	0.44 (0.39, 0.48)	0.45 (0.41, 0.48)	0.44 (0.39, 0.48)
σ^2	–	0.17 (0.09, 0.33)	0.14 (0.08, 0.24)	0.28 (0.16, 0.48)	1.14 (1.00, 1.32)	0.65 (0.46, 0.79)
τ^2	1.01 (0.97, 1.04)	0.93 (0.84, 0.98)	0.96 (0.87, 1.00)	0.74 (0.53, 0.86)	0.13 (0.08, 0.22)	0.52 (0.40, 0.64)
ϕ		0.016 (0.009, 0.051)	0.011 (0.007, 0.062)	0.016 (0.011, 0.023)	0.042 (0.031, 0.052)	0.05 (0.03, 0.065)
G	5935	5804	5759	5640	5751	5902
P	5939	5842	5830	5810	5777	5791
D	11874	11646	11590	11450	11529	11693
MSPE	2.00	1.98	1.96	1.94	1.96	1.96
Run time (hours)	–	9.44	19.31	37.08	19.42	9.73