An Additive Gaussian Process Approximation for Large Spatio-Temporal Data

Pulong Ma
Statistical and Applied Mathematical Sciences Institute
and Duke University
Durham, NC 27709

Bledar Konomi and Emily L. Kang
Department of Mathematical Sciences, University of Cincinnati

Abstract

We propose a new additive spatio-temporal Gaussian process approximation to model complex dependence structures for large spatio-temporal data. The proposed approximation method incorporates a computational-complexity-reduction method and a separable covariant function, which can capture different scales of variation and spatio-temporal interactions. The first component is able to capture nonseparable variation while the second component captures the separable variation of all scales. The Bayesian inference is carried out through a Markov chain Monte Carlo algorithm based on the hierarchical representation of the model. Through this representation we are able to utilize the computational advantages of both components. To demonstrate the inferential and computational benefits of the proposed method, we carry out four different simulation studies as well as a real application that concerns the spatio-temporal measurements of ground-level ozone in the Eastern United States.

Keywords: Additive model; Bayesian inference; Gaussian process; Metropolis-within-Gibbs sampler; nonseparable covariance function; spatio-temporal data
1 Introduction

Statistical modeling for spatio-temporal data can be proceeded in various ways depending on how the time domain is treated. We focus on a paradigm in which the space and time are both continuously indexed. One task in this avenue is to develop a class of computationally tractable nonseparable space-time covariance functions that capture different degrees of spatio-temporal interactions, since many geophysical processes evolve in space and time and they interact with each other. There is vast literature on constructing nonseparable space-time covariance functions (e.g., Cressie and Huang, 1999; De Iaco et al., 2002; Gneiting, 2002; Stein, 2005; Fonseca and Steel, 2011). Despite their ability to model spatio-temporal interactions, such nonseparable space-time covariance functions are generally computationally prohibitive for large spatio-temporal datasets due to $O(n^3)$ computational cost to calculate the Cholesky decomposition for the associated covariance matrix of $n$ data points and $O(n^2)$ memory cost.

Instead of devising a new class of nonseparable covariance functions, we offer another perspective to model spatio-temporal interactions by combining a nonseparable spatio-temporal process with a separable spatio-temporal process. The resulting spatio-temporal process can capture the spatio-temporal interactions through their variance components. To allow computational efficiency for the nonseparable spatio-temporal process, we will employ an approximation for the nonseparable spatio-temporal process.

Recently, various methods have been proposed to analyze large or massive datasets, including low-rank methods (Banerjee et al., 2008; Cressie and Johannesson, 2008; Finley et al., 2009), approximate likelihood methods (e.g., Stein et al., 2004; Gramacy and Apley, 2015), Gaussian Markov random fields (Lindgren et al., 2011), multiresolutional models (Nychka et al., 2013; Katzfuss, 2017), full-scale approximation (Sang and Huang, 2012; Zhang et al., 2018), nearest neighbor Gaussian process (Datta et al., 2016a), and a general framework for Vecchia approximations (Katzfuss and Guinness, 2017). These methods take advantages of a low-rank model, a low-order
conditioning set, or assumptions of a sparse covariance or precision matrix to alleviate the computational difficulty of GP modeling. Several previously mentioned computational-complexity-reduction methods for spatial data have been extended in the spatio-temporal context. For example, full-scale approximation (Sang and Huang, 2012) is extended to analyze spatio-temporal dataset (Zhang et al., 2015b) and multivariate computer model outputs (Zhang et al., 2015a). The nearest neighbor Gaussian process (NNGP; Datta et al., 2016a) is extended to analyze spatio-temporal dataset in Datta et al. (2016b). These methods pre-specify a parametric nonseparable space-time covariance function, and use various techniques to efficiently approximate the resulting GP with this covariance function. When these computational-complexity-reduction methods are applied to analyze spatio-temporal data, it is unclear whether these methods can offer a good approximation for a nonseparable covariance function while still maintaining computational efficiency. For instance, Datta et al. (2016b) fix the space-time interaction parameter in Gneiting’s nonseparable covariance function (Gneiting, 2002), but rather estimate it within the Markov chain Monte Carlo algorithm, which is a rather strong assumption to model spatio-temporal interactions. In environmental science, it is crucial to model the interaction among space and time explicitly for many geophysical processes (Cressie and Wikle, 2011).

We propose a spatio-temporal Gaussian process whose covariance function is constructed with an additive form of two components. The first component is nonseparable and constructed using a computational-complexity-reduction method aforementioned, while the second component assumes a separable covariance structure that increases the modeling flexibility and describes the potential spatio-temporal separability. We call the resulting method the additive approximate Gaussian process (AAGP), since its additive form of covariance functions can approximate any type of target covariance function, separable or nonseparable, in terms of its implied dependence structure, in a data-driven way. On the methodological end, although our method, AAGP, inherits the additive modeling form as some other methods such as the modified predictive process
(MPP; [Finley et al. 2009]), the full scale approximation (FSA; [Sang and Huang 2012]), and the multi-resolution approximation (MRA; [Katzfuss 2017]), our method differs from these methods substantially in the sense that although these methods use multiple components to approximate any parametric covariance function, this prespecified parametric covariance function is typically chosen to be a single covariance function based on exploratory analysis such as fitting empirical variogram models. Rather, our method uses two different covariance functions to model different dependence structures, where only the nonseparable covariance function is approximated with a computational-complexity-reduction method and the separable covariance function is not imposed with any approximation. As demonstrated in numerical studies, our method is accurate and flexible since it allows data to determine how much variation is explained by these two components, respectively, instead of simply assuming a single parametric covariance function that is either separable or nonseparable.

Our method also differs from previous methods using an additive structure resulting from two different covariance components. [Rougier 2008] uses a low-rank component plus a separable covariance function while the low-rank component is constructed with pre-specified regressors of input/output variables in a separable form. Our method explicitly includes the nonseparable dependence structure into the model which is not necessarily low-rank. In addition, our model includes a nugget term recommended in modeling environmental data ([Cressie 1993]), to ensure computational stability and better predictive performances. [Ma and Kang 2017] propose a model with a low-rank component and a Gaussian graphical model that induces a sparse precision matrix, but their method applies to spatial data. One more example for an additive covariance function model is the work in [Ba and Joseph 2012]. They model a process as a sum of two independent GPs with separable squared exponential covariance functions, and have to impose empirical constraints on parameters in these two covariance functions to avoid non-identifiability. Our method avoids such an issue by using different types of covariance structures for these two components and is
designed to handle large datasets in a Bayesian framework.

Gaussian process regression is usually implemented based on conventional likelihood-based inference or fully Bayesian inference, which typically focus on the marginalized model after integrating out random effects in the GPs (e.g., Ba and Joseph [2012], Banerjee et al. [2014]). Such inference procedures usually rely on the Sherman-Morrison-Woodbury formula (Henderson and Searle, 1981) or the Cholesky decomposition of a sparse matrix to reduce computational complexity when large amount of data is available (e.g., Sang and Huang [2012], Datta et al. [2016a], Katzfuss [2017]). However, these techniques cannot alleviate any computational burden in AAGP, since direct inference from the marginalized model is computationally infeasible for AAGP. To tackle this computational challenge, we propose a fully conditional Markov chain Monte Carlo (MCMC) algorithm to enable fast Bayesian inference in a general framework where a covariance function admits an additive form of multiple components. AAGP relies on a computational-complexity-reduction covariance function and a separable covariance function, and falls into this general framework. To elaborate the proposed general inference framework, the MPP model is chosen in AAGP to derive the computational-complexity-reduction covariance function. We will also show this general Bayesian inference still applies when other computational-complexity-reduction methods such as FSA, NNGP, and MRA are used.

The reminder of this paper is organized as follows. Section 2 presents the basic definition of the additive approximate Gaussian process and its covariance function specification. Section 3 gives details about the fast Metropolis-within-Gibbs sampler based on the fully hierarchical formulation of AAGP, and its extension with other covariance specifications in AAGP. Section 4 illustrates the predictive performance of AAGP with simulation examples and real data analysis. Section 5 concludes with discussions on possible extensions.
2 Additive approximate Gaussian process

Let \( \{Z(x) : x \in \mathcal{X} \equiv S \times T\} \) be a continuously-indexed spatio-temporal process, where \( x \equiv (s, u) \) with \( s \in S \) and \( u \in T \). \( S \subset \mathbb{R}^d \) is a \( d \)-dimensional spatial domain with positive integer \( d \), and \( T \subset \mathbb{R} \) is a temporal domain. Suppose that the spatio-temporal process \( Z(\cdot) \) is observed at a total of \( n \) locations, \( x_1, \ldots, x_n \in \mathcal{X} \). We assume the following model for \( Z(\cdot) \):

\[
Z(x) = Y(x) + \epsilon(x), \ x \in \mathcal{X},
\]

where \( Y(\cdot) \) is a latent Gaussian process of interest. The second term in the right-hand side is assumed to be a Gaussian white-noise process with variance \( \tau^2 \), which is usually called the nugget effect. This term is commonly used to represent measurement errors for environmental data.

The process \( Y(\cdot) \) is usually assumed to have additive components:

\[
Y(x) = h(x)^T b + w(x), \ x \in \mathcal{X},
\]

where \( h(\cdot) = [h_1(\cdot), h_1(\cdot), \ldots, h_p(\cdot)]^T \) is a vector of \( p \) covariates; \( b \) is the corresponding vector of \( p \) regression coefficients; \( w(\cdot) \) is a Gaussian process with mean zero and covariance function \( C_1(\cdot, \cdot) \). To reduce the computational complexity but also to increase flexibility in the dependence structure we assume that the process \( w(\cdot) \) is approximated with the summation of two computationally efficient components \( w_1(\cdot) \) and \( w_2(\cdot) \). We assume that \( w_1(\cdot) \) and \( w_2(\cdot) \) are independent Gaussian processes and have two different covariance functions families, \( C_1(\cdot, \cdot) \) and \( C_2(\cdot, \cdot) \) respectively. We call the resulting process \( Y(\cdot) \) the additive approximate Gaussian process (AAGP).

Its covariance function can be written as \( \text{cov}(Y(x), Y(x')) = C_1(x, x') + C_2(x, x') \) with the two components \( C_1(\cdot, \cdot) \) and \( C_2(\cdot, \cdot) \) described below.

In this paper we concentrate on two specific forms of the covariance functions. We choose a nonseparable covariance function \( C_1(\cdot, \cdot) \) to model spatio-temporal interactions. However, direct implementation of this covariance function is computationally challenging for large spatio-
temporal datasets. So, a computational-complexity-reduction method is used to approximate $C_1(\cdot, \cdot)$. Without loss of generality, the modified predictive process model is chosen to approximate $C_1(\cdot, \cdot)$ from several computational-complexity-reduction methods aforementioned. The covariance function $C_2(\cdot, \cdot)$ is chosen to be a separable covariance function. Since it is generally unclear how good the approximation is when a computational-complexity-reduction method is used to approximate a nonseparable covariance function, a separable covariance function can offer a tool to capture the lost information due to the usage of a computational-complexity-reduction method alone. In addition, the choice of a nonseparable covariance function avoids the non-identifiability issue since these two covariance functions characterize different dependence structures. Therefore, unlike imposing empirically strict constraints on parameters in [Ba and Joseph, 2012], our model can be fit in a purely data-driven way. As demonstrated by numerical examples in Section [H], with both of the computationally efficient nonseparable and separable components, AAGP is more flexible to model various dependence structures and gives robust predictive performance. Finally, AAGP consists of two different covariance components, which is fundamentally different from methods such as FSA and MRA in essence, because those methods are designed to use multiple components altogether to approximate a single target covariance function. In fact, both FSA and MRA are alternative computational-complexity-reduction methods to model the component $w_1(\cdot)$ in AAGP.

2.1 A computational-complexity-reduction covariance function

To deal with large data size $n$, we adopt an approximation method to reduce computational complexity resulting from $w_1(\cdot)$ with the nonseparable covariance function $C_1(\cdot, \cdot)$. Predictive process methods (e.g., [Banerjee et al., 2008, Finley et al., 2009]) have been proposed and applied successfully with large data. These methods assume models with reduced dimension and require only linear computational cost to invert large covariance matrix via the Sherman-Morrison-Woodbury formula. Specifically, we assume a nonseparable correlation function $R_0(\cdot, \cdot; \theta_1)$ known up to a
few parameters $\theta_1$. Pre-specifying a set of $m$ ($m \ll n$) knots $\mathcal{X}^* \equiv \{x_1^*, \ldots, x_m^*\} \subset \mathcal{X}$, we model the process $w_1(\cdot)$ as a GP with mean zero and correlation function given by:

$$R_1(x, x') = R(x, \mathcal{X}^*)R_*^{-1}R(x', \mathcal{X}^*)^T + I(x = x')[1 - R(x, \mathcal{X}^*)R_*^{-1}R(x', \mathcal{X}^*)^T],$$

where $R(x, \mathcal{X}^*) \equiv [R_0(x_i, x_j^*)]_{i=1,\ldots,m}$ is an $m$-dimensional row vector; $R_*$ is the $m$-by-$m$ matrix with its $(i,j)$-th element $R_0(x_i^*, x_j^*)$, for $i, j = 1, \ldots, m$, and $I(\cdot)$ denotes the indicator function of its argument. It is straightforward to show that $R_1(x, x') = 1$ if $x = x'$. Based on this construction, the correlation matrix of $w_1 \equiv (w_1(x_1), \ldots, w_1(x_n))^T$ is $R_1 \equiv R_{nm}R_*^{-1}R_{nm}^T + V$, where $R_{nm} \equiv [R_0(x_i, x_j^*)]_{i=1,\ldots,n;j=1,\ldots,m}$, and $V$ is an $n$-by-$n$ diagonal matrix with its $i$th diagonal element given by $V_i \equiv 1 - R(x_i, \mathcal{X}^*)R_*^{-1}R(x_i, \mathcal{X}^*)^T$, for $i = 1, \ldots, n$. Note that the vector $R(x, \mathcal{X}^*)$ and matrices $R_*$ and $V$ all depend on the unknown parameters $\theta_1$. The resulting covariance function $C_1(\cdot, \cdot; \theta_1)$ is $\sigma_1^2 R_1(\cdot, \cdot; \theta_1)$, where $\sigma_1^2$ is a variance parameter. Readers are referred to [Finley et al. 2009] for more details on MPP. Although MPP is used to demonstrate AAGP as one instance of computational-complexity-reduction methods, we discuss in Section 3 on how the general inference framework will apply if other computational-complexity-reduction methods are used.

### 2.2 A separable covariance function

We assume a separable covariance function for the process $w_2(\cdot)$. For $s, s' \in S, u, u' \in \mathcal{T}$, the process $w_2(\cdot)$ is thus assumed to have variance $\sigma^2$ with the following separable correlation function

$$R_2(x, x'; \theta_2) = \rho_1(s, s'; \phi_1)\rho_2(u, u'; \phi_2),$$

where $\rho_1(\cdot, \cdot)$ and $\rho_2(\cdot, \cdot)$ are correlation functions with range parameters $\phi_1$ and $\phi_2$ over space $S$ and $\mathcal{T}$, respectively. $\theta_2 \equiv \{\phi_1, \phi_2\}$ denotes these two range parameters. For the present, the response $Z(\cdot)$ is assumed to be observed at all the $n = n_1n_2$ locations arranged as $x_1 = \ldots = x_n = \ldots = x_{n_2} = \ldots$
\((s_1, u_1), \ldots, x_{n_2} = (s_1, u_{n_2}), x_{n_2+1} = (s_2, u_1), \ldots, x_{2n_2} = (s_2, u_{n_2}), \ldots, x_n = (s_{n_1}, u_{n_2})\), where \(n_1\) denotes the number of locations in \(S\), and \(n_2\) denotes that in \(T\). The locations \(\{s_1, \ldots, s_{n_1}\}\) and \(\{u_1, \ldots, u_{n_2}\}\) are not necessarily regularly spread out in \(S\) and \(T\). The resulting correlation matrix of \(w_2 \equiv (w_2(x_1), \ldots, w_2(x_n))^T\) is \(R_2 \equiv R_s \otimes R_u\), where \(R_s \equiv [\rho_1(s_i, s_j)]_{i,j=1,\ldots,n_1}\) is an \(n_1\)-by-\(n_1\) matrix, and \(R_u \equiv [\rho_2(u_i, u_j)]_{i,j=1,\ldots,n_2}\) is an \(n_2\)-by-\(n_2\) matrix. As shown in Genton (2007) and Rougier (2008), imposing separability on the covariance function enables us to use attractive properties of Kronecker product of matrices, which brings substantial computational gains to calculations involving Kronecker product. In addition, we focus on the problem that \(n\) is large (in order of \(10^4 \sim 10^6\)) but \(n_1\) and \(n_2\) are small (less than \(10^3\)). In Section 5, several modeling strategies are discussed when either \(n_1\) or \(n_2\) is large. The tentative assumption that \(Z(\cdot)\) is observed at all the \(n = n_1n_2\) locations will be relaxed in Section 3.4. We will illustrate there how missing data imputation step can be embedded in the Bayesian inference.

### 2.3 Likelihood evaluation

Let \(Z \equiv (Z(x_1), \ldots, Z(x_n))^T\) be the \(n\)-dimensional vector of observations. Given the model and the covariance structure, as specified above, the log-likelihood function of the data vector \(Z\) can be written as

\[
\ell(b, \tau^2, \sigma_1^2, \sigma_2^2, \theta_1, \theta_2; Z) = -n \log(2\pi)/2 - (Z - Hb)^T \Sigma^{-1}(Z - Hb)/2,
\]

where \(H \equiv [h(x_1), \ldots, h(x_n)]^T\) is a matrix of covariates or regressors. \(\Sigma\) is the covariance matrix of \(Z\) with the following form

\[
\Sigma \equiv \text{cov}(Z) = \sigma_1^2 R_1 + \sigma_2^2 R_s \otimes R_u + \tau^2 I,
\]

\[
= \sigma_1^2 R_{nm} R_s^{-1} R_{nm}^T + \sigma_2^2 R_s \otimes R_u + \sigma_1^2 V + \tau^2 I.
\]
Evaluation of this log-likelihood function involves the inversion and determinant of the $n$-by-$n$ covariance matrix $\Sigma$. When $n$ is large, the Sherman-Morrison-Woodbury formula or the Cholesky decomposition of sparse matrices have been suggested to reduce computational complexity (e.g., [Sang and Huang, 2012], [Datta et al., 2016a], [Katzfuss, 2017]). However, these techniques cannot reduce computational complexity for the likelihood evaluation in AAGP and calculations related to the covariance matrix $\Sigma$. Specifically, let the symbol $D$ denote the matrix $\sigma_2^2R_s \otimes R_u + \sigma_1^2V + \tau^2I$ to simplify notations. Then the Sherman-Morrison-Woodbury formula can be applied to derive the formula for $\Sigma^{-1}$,

$$
\Sigma^{-1} = D^{-1} - D^{-1}R_{nm}[R_s/\sigma_1^2 + R_{nm}^TD^{-1}R_{nm}]^{-1}R_{nm}^TD^{-1},
$$

where the inversion of $D$ is required in order to solve linear systems involving $\Sigma$. It should be noted that calculating this inversion $\Sigma^{-1}$ is not computationally feasible for large $n$. In particular, it requires inversions of two $m$-by-$m$ matrices, $R_s$ and $R_s/\sigma_1^2 + R_{nm}D^{-1}R_{nm}$, and inversion of the $n$-by-$n$ matrix $D$. As $m$ is much smaller than $n$, inverting the $m$-by-$m$ matrices can be done easily with $O(m^3)$ flops since we have $m$ much smaller than $n$. However, inverting the $n$-by-$n$ matrix $D$ requires full matrix inversion due to the presence of heterogeneous diagonal elements in $\sigma_1^2V$, and thus requires $O(n^3)$ flops and $O(n^2)$ memory. Classical kriging approach and Bayesian inference based on the marginalized form of the likelihood in (2.3) is hence computationally too expensive. To overcome this difficulty, a fully conditional MCMC algorithm based on the hierarchical representation of the model is proposed in Section 3.1 to allow fast Bayesian inference.

### 3 Bayesian inference: A fully conditional approach

To carry out Bayesian inference for AAGP, we first assign prior distributions to the unknown parameters $\{b, \tau^2, \sigma_1^2, \sigma_2^2, \theta_1, \theta_2\}$. Following customary prior specifications, we assign a vague multivariate normal prior to the coefficient vector $b \sim N_p(\mu_b, V_b)$; inverse gamma priors are
assigned to variance parameters, \(\tau^2 \sim \mathcal{IG}(a_\tau, b_\tau)\), \(\sigma_1^2 \sim \mathcal{IG}(a_1, b_1)\), and \(\sigma_2^2 \sim \mathcal{IG}(a_2, b_2)\); other parameters in \(\theta_1\) and \(\theta_2\) such as range parameters are assigned with uniform priors.

Conventional fully Bayesian inference procedures for GP modeling typically focus the marginalized model after integrating out random effects, and obtain posterior distribution of unknown parameters given data. For AAGP, we can write out the (joint) posterior distribution \(p(b, \tau^2, \sigma_1^2, \sigma_2^2, \theta_1, \theta_2 \mid Z)\), which is proportional to the joint distribution:

\[
p(b, \tau^2, \sigma_1^2, \sigma_2^2, \theta_1, \theta_2) p(Z \mid b, \tau^2, \sigma_1^2, \sigma_2^2, \theta_1, \theta_2) \propto N_p(\mu_b, V_b) \mathcal{IG}(a_\tau, b_\tau) \mathcal{IG}(a_1, b_1) \mathcal{IG}(a_2, b_2) p(\theta_1, \theta_2) \times N_n(Hb, \Sigma).
\]

Sampling from this posterior distribution (3.1) is computationally infeasible with large \(n\), since each MCMC iteration requires inversion of the \(n\)-by-\(n\) covariance matrix \(\Sigma\) and hence costs \(O(n^3)\) flops and \(O(n^2)\) memory to compute the likelihood. Here, rather than utilizing the marginal distribution of \(Z\), we write the model in a hierarchical form of AAGP involving the latent processes \(w_1(\cdot)\) and \(w_2(\cdot)\), based on which we propose a computationally efficient MCMC sampler.

The models (2.1) and (2.2) can be written as follows to give a hierarchical formulation of AAGP:

\[
Z \mid b, w_1, w_2, \tau^2 \sim N_n(Hb + w_1 + w_2, \tau^2 I_n),
\]

\[
w_1 \mid \sigma_1^2, \theta_1 \sim N_n(R_{nm} R_x^{-1} w^*, \sigma_1^2 V),
\]

\[
w_2 \mid \sigma_2^2, \theta_2 \sim N_n(0, \sigma_2^2 R_s \otimes R_u),
\]

where \(I_n\) denotes the \(n\)-by-\(n\) identity matrix, and \(w^* \equiv (w_1(x_1^*), \ldots, w_1(x_m^*))^T\) is a vector of length \(m\) following multivariate normal distribution with mean zero and covariance matrix \(\sigma_1^2 R_s\). The joint posterior distribution of all the unknowns, including parameters \(\{b, \tau^2, \sigma_1^2, \sigma_2^2, \theta_1, \theta_2\}\)
and latent random effects $w^*, w_1,$ and $w_2,$ can be obtained:

$$
p(b, \tau^2, \sigma_1^2, \sigma_2^2, \theta_1, \theta_2, w^*, w_1, w_2 \mid Z)$$

$$\propto N_p(\mu_b^*, V_b^*)IG(a_r, b_r)IG(a_1, b_1)IG(a_2, b_2)p(\theta_1, \theta_2)$$

$$\times N_m(w^* \mid 0, \sigma_2^2 R_s) \times N_n(w_1 \mid R_{nm} R_s^{-1} w^*, \sigma_1^2 V)$$

$$\times N_n(w_2 \mid 0, \sigma_2^2 R_s \otimes R_a) \times N_n(Z \mid Hb + w_1 + w_2, \tau^2 I_n).$$

### 3.1 Parameter estimation & computational cost

Since the posterior distribution (3.5) is intractable, we use a Metropolis-within-Gibbs sampler (Hastings, 1970; Gelfand and Smith, 1990) for parameter inference in the hierarchical representation of the posterior distribution. In particular, the conjugate full conditional distributions for $b, \tau^2, \sigma_1^2, \sigma_2^2,$ and multivariate normal full conditional distributions for random effects $w^*, w_1,$ and $w_2$ are given below:

- $b \mid \cdot \sim N_p(\mu_b^*, V_b^*),$
- $\sigma_1^2 \mid \cdot \sim IG\left(a_1 + \frac{m + n}{2}, b_1 + \frac{w^* R_s^{-1} w^* + (w_1 - R_{nm} R_s^{-1} w^*)^T V^{-1} (w_1 - R_{nm} R_s^{-1} w^*)}{2}\right),$
- $\sigma_2^2 \mid \cdot \sim IG\left(a_2 + \frac{m + n}{2}, b_2 + \frac{w_2^T (R_s \otimes R_a)^{-1} w_2}{2}\right),$
- $\tau^2 \mid \cdot \sim IG\left(a_r + \frac{m}{2}, b_r + \frac{(Z - Hb - w_1 - w_2)^T (Z - Hb - w_1 - w_2)}{2}\right),$
- $w^* \mid \cdot \sim N_m(\mu_{w^*}, \Sigma_{w^*}),$
- $w_1 \mid \cdot \sim N_n(\mu_{1|}, \Sigma_{1|}),$
- $w_2 \mid \cdot \sim N_n(\mu_{2|}, \Sigma_{2|}),$

where $\mu_b^* = V_b^* [\tau^{-2} H^T (Z - w_1 - w_2) + V_b^{-1} \mu_b],$ and $V_b^* = [\tau^{-2} H^T H + V_b^{-1} ]^{-1};$ $\mu_{w^*} = \Sigma_{w^*} R_s^{-1} R_{nm} (\sigma_2^2 V)^{-1} w_1,$ and $\Sigma_{w^*} = \sigma_2^2 R_s (R_s + R_{nm} V^{-1} R_{nm})^{-1} R_s^T;$ $\mu_{1|} = \Sigma_{1|} [(\sigma_2^2 V)^{-1} R_{nm} R_s^{-1} w^* + \tau^{-2} (Z - Hb - w_2)],$ and $\Sigma_{1|} = [\tau^{-2} I_n + (\sigma_2^2 V)^{-1} ]^{-1};$ $\mu_{2|} = \tau^{-2} \Sigma_{2|} (Z - Hb - w_1).$
and \( \Sigma_{2|} = [\tau^{-2}I_n + (\sigma_2^2 R_s \otimes R_u)^{-1}]^{-1} \).

Close inspection of these full conditional distributions \( p(w^* | \cdot), p(w_1 | \cdot), p(w_2 | \cdot), p(b | \cdot), p(\sigma_1^2 | \cdot), p(\sigma_2^2 | \cdot), p(\tau^2 | \cdot) \) reveals that sampling from them only requires inversions of \( n \times n \) diagonal matrices, \( m \times m \) low-dimensional matrices, and \( n_1 \times n_1, n_2 \times n_2 \) small matrices, thus making this inference procedure computationally attractive. For instance, to sample from the distribution \( [w_2 | \cdot] = \mathcal{N}_n(\mu_2, \Sigma_{2|}) \), we can obtain the eigenvalue decomposition of \( \Sigma_{2|} \) using the properties of Kronecker product of two matrices. Specifically, we first carry out eigenvalue decomposition for matrices \( R_s = u_1 \Lambda_1 u_1^T \) and \( R_u = u_2 \Lambda_2 u_2^T \), which require computational cost \( O(n_1^3 + n_2^3) \) in total. Then we generate an \( n \)-dimensional vector of independent samples from the standard normal distribution, denote by \( \zeta \). The sample from \( \mathcal{N}_n(\mu_2, \Sigma_{2|}) \) is obtained by carrying out the matrix-vector multiplication, \( (u_1 \otimes u_2)\Lambda^{1/2}\zeta \), where \( \Lambda^{1/2} \) denotes the square root of the diagonal matrix \( (\Lambda_1^{-1} \otimes \Lambda_2^{-1}/\sigma_2^2 + \tau^{-2}I_n)^{-1} \). This matrix-vector multiplication requires computational cost \( O(n_1^3 + n_2^3 + n(n_1 + n_2)) \).

To sample \( \theta_1 \) and \( \theta_2 \) from their full conditional distributions, a Metropolis-Hastings step is incorporated for each parameter. These full conditional distributions are not any standard distribution, but we know them up to some normalizing constant as follows:

\[
p(\theta_1 | \cdot) \propto p(\theta_1) \times \mathcal{N}_m(w^* | 0, \sigma_1^2 R_s) \times \mathcal{N}_n(w_1 | R_{nm} R_s^{-1} w^* , \sigma_1^2 V) \tag{3.6}
\]

\[
p(\theta_2 | \cdot) \propto p(\theta_2) \times \mathcal{N}_n(w_2 | 0, \sigma_2^2 R_s \otimes R_u) \tag{3.7}
\]

Let \( \theta \) denote a generic scalar parameter in \( \theta_1 \) or \( \theta_2 \) with the target probability density \( \pi(\theta) \) given in Eq. (3.6) or Eq. (3.7). Let \( J(\theta^* | \theta) \) denote the proposal distribution for \( \theta \). Then the corresponding acceptance ratio \( \alpha_0 \) in the Metropolis-Hastings step for \( \theta \) is

\[
\alpha_0 = \min \left\{ 1, \frac{\pi(\theta^*)J(\theta^* | \theta)}{\pi(\theta)J(\theta | \theta^*)} \right\}.
\]

Let \( \mu \) be a random number generated from \( U([0, 1]) \). The proposed candidate \( \theta^* \) is accepted if \( \alpha_0 > \mu \); otherwise, the proposed candidate is rejected.
The Metropolis-within-Gibbs sampler described above is computationally efficient. In terms of computational cost, sampling from the full conditional distributions of \( b, \tau^2, \sigma_1^2, \sigma_2^2, w^*, w_1 \), and \( w_2 \) requires \( O(m^3 + m^2n + n_1^3 + n_2^3 + n(n_1 + n_2)) \) flops. Sampling from full conditional distributions for \( \theta_1 \) and \( \theta_2 \) requires \( O(m^3 + m^2n + n_1^3 + n_2^3 + n(n_1 + n_2)) \) flops. Therefore, the overall computational cost for each MCMC iteration is \( O(m^2n + n_1^3 + n_2^3 + n(n_1 + n_2)) \). Note that \( m, n_1, \) and \( n_2 \) are all smaller than \( n \), which makes this inference procedure much more efficient than dealing with marginalized joint density by integrating random effects. It is worth mentioning that if either \( n_1 \) or \( n_2 \) is large, it is desirable or even necessary to further reduce computational complexity of the current Bayesian inference procedure, and we discuss in Section 5 on how to extend AAGP and the Metropolis-within-Gibbs sampler for very large data with large \( n_1 \) or \( n_2 \). Moreover, although we sample the \( n \)-dimensional vectors \( w_1 \) and \( w_2 \) in the Gibbs sampler, it is worth noting that there is no need to store all samples of these two high-dimensional vectors from all MCMC iterations, because they can always be recovered easily through 
\[
\begin{align*}
[w_1|Z] &= \int [w_1|\sigma_1^2, \theta_1][\sigma_1^2, \theta_1|Z] d\{\sigma_1^2, \theta_1\} \\
[w_2|Z] &= \int [w_2|\sigma_2^2, \theta_2][\sigma_2^2, \theta_2|Z] d\{\sigma_2^2, \theta_2\}.
\end{align*}
\]
Therefore, the overall memory cost for each MCMC iteration is roughly \( O(mn + n_1^3 + n_2^3) \).

To further improve computational efficiency, the random vector \( w^* \) and parameter \( \sigma_1^2 \) can be integrated out when sampling from \( p(\theta_1 | \cdot) \equiv p(\theta_1 | w_1, w^*, \sigma_1^2) \). That is, we seek the joint conditional distribution \( p(\theta_1, w^*, \sigma_1^2 | w_1) \), and then integrate out \( w^*, \sigma_1^2 \) to obtain the conditional distribution of \( \theta_1 \) given \( w_1 \). Similar efficiency can be achieved to obtain samples of \( \theta_2 \) from \( p(\theta_2 | \cdot) \equiv p(\theta_2 | w_2, \sigma_2^2) \). That is, we seek the joint conditional distribution \( p(\theta_2, \sigma_2^2 | w_2) \), and then integrate out \( \sigma_2^2 \).

### 3.2 Prediction

For any location \( x_0 = (s_0, u_0) \in \mathcal{X} \), our interest is to make prediction for \( Y(x_0) \). Define \( \Omega = \{b, \sigma_1^2, \sigma_2^2, \tau^2, \theta_1, \theta_2\} \). The (posterior) predictive distribution of \( Y(x_0) \) given \( Z \) is derived
as follows:

\[
p(Y(x_0) | Z) = \int p(Y(x_0) | w_1, w_2, \Omega, Z)p(w_1, w_2, \Omega | Z) \, dw_1 dw_2 d\Omega
\]

\[
= \int p(Y(x_0) | w_1, w_2, \Omega)p(w_1, w_2, \Omega | Z) \, dw_1 dw_2 d\Omega.
\]

Here, \( p(Y(x_0) | w_1, w_2, \Omega) \) is the multivariate normal with mean \( \hat{Y}(x_0) \equiv h(x_0)^T b + \hat{w}_1(x_0) + \hat{w}_2(x_0) \), and variance \( \hat{\sigma}^2(x_0) = \hat{\sigma}_1^2(x_0) + \hat{\sigma}_2^2(x_0) \), which are

\[
\hat{w}_1(x_0) = [R_{nm} R_s^{-1} R(x_0, \mathcal{X}^*)]^{T} R_1^{-1} w_1,
\]

\[
\hat{w}_2(x_0) = (\rho_{1,0} \otimes \rho_{2,0})^{T} (R_s \otimes R_u)^{-1} w_2,
\]

\[
\hat{\sigma}_1^2(x_0) = \sigma_1^2 - \sigma_1^2 [R_{nm} R_s^{-1} R(x_0, \mathcal{X}^*)]^{T} R_1^{-1} R_{nm} R_s^{-1} R(x_0, \mathcal{X}^*)^{T},
\]

\[
\hat{\sigma}_2^2(x_0) = \sigma_2^2 - \sigma_2^2 (\rho_{1,0} \otimes \rho_{2,0})^{T} (R_s \otimes R_u)^{-1} (\rho_{1,0} \otimes \rho_{2,0}),
\]

with \( \rho_{1,0} = (\rho_1(s_0, s_1), \ldots, \rho_1(s_0, s_{n_1}))^{T} \) and \( \rho_{2,0} = (\rho_2(u_0, u_1), \ldots, \rho_2(u_0, u_{n_2}))^{T} \).

Sampling from this predictive distribution involves inversion of matrices \( R_1 \) and \( R_s \otimes R_u \). The matrix \( R_1 \) can be inverted efficiently via the Sherman-Morrison-Woodbury formula

\[
R_1^{-1} = V^{-1} - V^{-1} R_{nm} (R_s \otimes R_u) R_{nm} V^{-1} R_{nm}^{-1} R_s^{-1} V^{-1},
\]

where the inversion of \( R_1 \) only requires computational cost \( O(nm^2) \). Using the following result:

\[
\text{vec}(T) = (R_s \otimes R_u)^{-1} \text{vec}(K) \Leftrightarrow T = R_u^{-1} K R_s^{-T},
\]

we can solve linear systems involving the Kronecker product of \( R_s \) and \( R_u \) efficiently. The overall computational cost is \( O(m^3 + m^2 n + n_1^3 + n_2^3 + n(n_1 + n_2)) \) for taking samples from the predictive distribution of \( Y(\cdot) \) give \( Z \). Samples from the predictive distribution \( p(Y(x_0) | Z) \) can be obtained using composition sampling technique. That is, we draw from \( p(Y(x_0) | w_1, w_2, \Omega) \), with posterior samples of \( w_1, w_2, \Omega \) plugged into this distribution.
3.3 Alternative specification

The general framework in AAGP relies on a computational-complexity-reduction covariance function model and a separable covariance function model. The modified predictive process is chosen to derive the computational-complexity-reduction covariance function and to illustrate the computational benefit of the proposed fast Bayesian inference procedure. As recently noted in Katzfuss and Guinness (2017), MPP is a special case of more general Vecchia approximations, which include other existing methods such as FSA, NNGP, and MRA. These methods can also be used to derive the computational-complexity-reduction covariance function, and our fully conditional Bayesian inference still works for large datasets. In particular, the matrix $V$ in Eq. (2.4) will be replaced by a sparse matrix resulted from tapering when FSA is used. The proposed inference procedure described above can be applied efficiently. For NNGP and MRA, the vector $w^*$ then will be high-dimensional, because these two methods using a smaller number of conditioning set to construct a sparse precision matrix rather than resorting to a low-rank structure for the covariance matrix. The resulting covariance matrix $R^{-1}$ of $w^*$ is a sparse matrix. Therefore, our inference procedure can be implemented efficiently as well.

3.4 Missing data imputation

Recall that we represent $\mathcal{X}$ as a product space $S \times T$ and have tentatively assumed that the response $Z(\cdot)$ is observed at all $n = n_1n_2$ locations, where $n_1$ denotes the number of locations in $S$, and $n_2$ denotes that in $T$. Now we relax this assumption and describe how missing data imputation can be carried out. Specifically, we introduce $D_c \equiv \{(s_i, u_j) : s_i \in S, u_j \in T, i = 1, \ldots, n_1; j = 1, \ldots, n_2\}$ to denote the complete grid over $\mathcal{X} = S \times T$. We assume that the response variable $Z(\cdot)$ is only observed at a subset of $n$ ($n < n_1n_2$) locations $D_o \equiv \{x_1, \ldots, x_n\} \subset D_c$. The resulting $n$-dimensional data vector is denoted by $Z_o$, and we let $Z_m$ denote the $(n_1n_2 - n)$-dimensional vector of $Z(\cdot)$ at the unobserved locations in $D_m \equiv D_c \setminus D_o$. In the Metropolis-within-Gibbs
sampler, we now use $w_1$, $w_2$, and $Z$ to represent the $(n_1n_2)$-dimensional vectors at all locations in $D_o$, and treat $Z_m$ as unknown. The full conditional distributions and sampling procedure for parameters $\{b, \tau^2, \sigma_1^2, \sigma_2^2, \theta_1, \theta_2\}$ and random effects $w^*$, $w_1$ and $w_2$ are the same as described in Section 3.1. The missing values $Z_m$ can also be easily updated in MCMC based on its full conditional distribution. Actually, it can be shown that $[Z_m \mid \cdot] = N(Y_m, \tau^2 I)$, where $Y_m \equiv H_m b + w_{1,m} + w_{2,m}$, and $H_m$, $w_{1,m}$, and $w_{2,m}$ denote the matrix of covariates and subsets of the random effects $w_1$ and $w_2$ over the unobserved locations in $D_m$, respectively.

4 Numerical illustrations

This section presents several simulation examples under different scenarios to illustrate the model accuracy and predictive accuracy. The proposed method additive approximate Gaussian process is compared with modified predictive process and nearest neighbor Gaussian process. In addition, the full Gaussian process, referred to as Full GP, is used as benchmark in the synthetic examples. The proposed method is also applied to analyze Eastern US ozone data. All the methods are implemented in MATLAB R2015b, and the algorithms are run in a 10-core HP Intel Xeon E5-2680 machine with 12 GB random-access memory.

Section 4.1 uses three artificial examples to show whether AAGP can offer any computational and inferential benefits over other methods such as MPP and NNGP when the underlying true fields show different types of spatio-temporal dependence structures. Section 4.2 demonstrates the performance of AAGP when the underlying true field is generated from a deterministic function. In the following numerical examples, Gneiting’s nonseparable correlation function (Gneiting, 2002) is used, since this correlation function is widely used to model the interaction among different variables and is easy to interpret, although other types of nonseparable correlation functions are available as well. In particular, the Gneiting’s nonseparable correlation function has the following
form
\[
\rho((s, u), (s', u')) = \left(\frac{(u - u')^{2\alpha}}{a} + 1\right)^{-d/2} \cdot \exp\left\{-\frac{\|s - s'\|}{c\left(\frac{(u - u')^{2\alpha}}{a} + 1\right)^{\beta/2}}\right\},
\]
(4.1)
where \(d\) is the dimension of the spatial domain \(S\); \(a\) is the temporal range parameter in \(T\); \(c\) is the spatial range parameter in \(S\); \(\alpha \in (0, 1]\) is the smoothness parameter in \(T\); \(\beta \in [0, 1]\) is the interaction parameter between \(S\) and \(T\).

To compare AAGP with other methods, we use posterior summary of 50 (2.5, 97.5) percentiles of model parameters, mean-squared-prediction errors (MSPEs), and average length of 95% credible intervals (ALCI) for predictive values, to assess model adequacy and predictive accuracy. In addition, we also report total computing time for each model.

4.1 Artificial examples with known covariance structures

To demonstrate the inferential and computational benefit of AAGP, three different scenarios with different space-time covariance structures will be implemented for the simulated true field \(Y(\cdot)\) in a spatio-temporal domain \(\mathcal{X} = [0, 20]^2 \times [1, 20]\). Specifically, the following three scenarios for the underlying true field \(Y(\cdot)\) are considered:

1. Gneiting’s space-time nonseparable correlation function only (First Scenario);
2. separable correlation function only (Second Scenario);
3. a combination of Gneiting’s space-time nonseparable correlation function and separable correlation function (Third Scenario).

The covariates in trend term contain \(h_1(x), h_2(x)\), where \(h_1(x)\) is simulated for standard normal distribution, and \(h_2(x) = \cos(1^T x)\) for \(x \in \mathcal{X}\). Then the true process \(Y(\cdot)\) is simulated on 225 randomly selected spatial locations in two dimensional space \(S = [0, 20] \times [0, 20]\) and 20 locations in time \(T = [1, 20]\), which results in 4500 locations in the spatio-temporal domain \(\mathcal{X}\). The data are obtained by adding measurement noise or nugget effect. For all the simulated data, 90% of them
are randomly selected as training set to carry out Bayesian inference, and the remaining 10% are used to compare the predictive performance.

In each scenario, four different models have been implemented: Full GP, MPP, NNGP and AAGP. The Full GP model is used as a benchmark. MPP and NNGP are two instances of computational-complexity-reduction methods. In practice, one typically prespecifies a single target covariance function in these computational-complexity-reduction methods, say Gneiting’s nonseparable covariance function in our discussion. In all the scenarios, we will use a single target covariance function in MPP and NNGP. As we only focus on MPP as a specific implementation of AAGP in our simulation study, MPP will be a sub-model of AAGP. Previous work has shown that NNGP gives better results than MPP \cite{Dattaetal2016} for spatial and spatio-temporal data. So, AAGP with MPP as the computational-complexity-reduction method can be a competing method over NNGP. Notice that although we use MPP in AAGP in our discussion, any other computational-complexity-reduction methods including NNGP can be incorporated in AAGP as well. Each model is configured as follows. The MPP model is implemented with 250 knots and 704 knots, respectively, where 704 knots are used so that MPP will take about the same amount of time as in AAGP. AAGP is implemented with just 250 knots. The NNGP model is implemented with 15 nearest neighbors using sequential update in the MCMC algorithm shown in \cite{Dattaetal2016}. For each model, independent customary prior distributions are assigned for all parameters: (1) $b \sim N_2(0, 1000I)$; (2) $\sigma^2_1 \sim IG(2, 0.01)$; (3) $a \sim U(0, 20)$; (4) $c \sim U(0, 20)$; (5) $\beta \sim U(0, 1)$; (6) $\tau^2 \sim IG(2, 0.01)$; The smoothness parameter $\alpha$ is fixed at 0.5. The prior distributions for parameters in the space-time separable covariance function are specified as: $\sigma^2_2 \sim IG(2, 0.01)$, $\phi_s \sim U(0, 20)$, $\phi_t \sim U(0, 20)$. The MCMC algorithm is run 25000 iterations for each model with a burn-in period of 15000 iterations indicating independence from convergence diagnostics. In addition, we also add very small fixed nuggets to spatial and temporal separable correlation matrices $R_s$ and $R_u$ to avoid numerical instabilities in the MCMC algorithm, since their inverses are required.
In the first scenario, the latent true process $Y(\cdot)$ is assumed to have a Gneiting’s space-time covariance function, and their parameter specifications are listed in the second column of Table [1]. The results obtained from these four different models are reported in Table [1]. AAGP gives better prediction results than MPP and NNGP, since the MSPE in AAGP is more than 20% smaller than that in MPP and NNGP, and is close to the MSPE in Full GP. This confirms that the separable component in AAGP can capture part of the unexplained variability from the MPP. The spread of predictive distribution is very similar for both MPP and AAGP, but the predictive distribution for AAGP is slightly accurate than that for MPP, and slightly worse than that for NNGP. In AAGP, the regression coefficients $b_1, b_2$ are estimated very well in comparison to the results in Full GP. The mean of the posterior samples for the variance parameters are $\sigma^2_1 = 0.869$ and $\sigma^2_2 = 0.303$ for the first and second component. AAGP shows a clear preference for the Gneiting’s nonseparable covariance function in this case. The proposed AAGP seems to be able to automatically assign the variation missed by the low-rank component to the separable component and the nugget. Despite the fact that the nugget is under-estimated and the overall variance $\sigma^2_1 + \sigma^2_2$ overestimated, the predictive performance of the proposed method is better. One possible explanation of the above value is that the sum of estimated $\sigma^2_2$ and $\tau^2$ both together play the role of “nugget effect”. It is worth noting that the space-time interaction parameter $\beta$ has very wide credible interval even in the Full GP model, which indicates that this parameter cannot be estimated accurately even under the true model.

In the second scenario, the latent true process $Y(\cdot)$ is assumed to have squared exponential correlation functions in space and time with parameters specified in the second column of Table [2]. Simulation results in Table [2] suggest that both MPP and NNGP fail to detect the separability of the true field with the Gneiting’s space-time correlation function, since $\beta$ has 95% credible interval spreading out almost its entire support $[0, 1]$. Compared to MPP, AAGP can detect the separability, and it gives much better prediction results. In addition, the variance parameter $\sigma^2_1$ is close to 0,
and variance parameter $\sigma^2_2$ is close to 1, which indicates that AAGP is able to select the most appropriate model for the data. Compared to NNGP, AAGP also gives better MSPE and ALCI. The prediction results in AAGP are very close to the results of Full GP, and much better than that obtained from MPP and NNGP. The trend parameters $b_1, b_2$ can also be estimated very well. We can see that the posterior mean of the variance parameter $\sigma_1^2$ is close to 0, and the posterior mean of the variance parameter $\sigma_2^2$ is close to 1. These two variance parameters serve as weights for the two components in AAGP, and they are correctly identified: $\sigma_1^2 = 0.03$ and $\sigma_2^2 = 0.996$. The results in MPP also show that disadvantage is obvious when MPP is used to approximate a covariance function that is not the same one implied by the data, as the estimate of nugget is even larger than the estimate of variance in MPP. The failure of MPP and NNGP on detecting the separability may also rely on the fact the Gneiting’s space-time correlation function is not as smooth as the process with the squared exponential correlation function. To improve results for MPP and NNGP, we tried to take $\alpha$ to be a random variable. However, this leads to computational instabilities in the MCMC algorithm.

In the third scenario, we addresses the problem of parameter estimation and predictive performance in AAGP under a true covariance function model. The process $Y(\cdot)$ is simulated from an additive Gaussian process with a Gneiting’s space-time covariance function and a separable squared exponential covariance function with parameters specified in the second column of Table 3. Simulation results for each model are reported in Table 3. Compared to MPP, the MSPE in AAGP is 66% smaller than that in MPP, and close to the MSPE in Full GP. The predictive performance in AAGP is relatively close to the predictive performance of Full GP. The posterior mean of $b_1, b_2, \beta, \sigma_2^2, \phi_s, \phi_t$ are well estimated in AAGP. The variance parameter $\sigma_1^2$ and range parameter $\alpha$ are slightly over-estimated. We also observe that the nugget is underestimated in AAGP. This is likely because there are fixed small constants $\tau_s^2, \tau_u^2$ added to the diagonal of the separable correlation matrices $R_s$ and $R_u$. As a consequence, it makes the nugget term to be $(\tau^2 + \tau_s^2 \tau_u^2)$ instead...
of $\tau^2$. As the variance parameters $\sigma_1^2$ and $\sigma_2^2$ are correctly identified in AAGP, the proposed model AAGP is able to correctly identify the variation coming from the non-separable and separable part. The range parameter $c$ in AAGP is over-estimated, and this is likely due to the overestimation of the variance parameter $\sigma_1^2$, since their ratio plays an important role in predictions under certain conditions (for details, see [Kaufman and Shaby, 2013]).

To briefly summarize our findings, in all the examples above AAGP can give better predictive performance than MPP even though more knots are included in MPP, regardless of the simulation settings. This proves that by adding an additional separable covariance function model, AAGP outperforms its computational-complexity-reduction method alone. In MPP, uniformly spaced knots are selected, and more sophisticated way to select the knots in MPP is beyond the scope of this paper, for details, see [Ghahaniyogi et al., 2011]. In addition, AAGP also gives better results than NNGP in terms of capturing spatio-temporal interactions. In the first scenario, we compared AAGP with its sub-model MPP and MPP with more knots. Even though the covariance function model in simulated true field favors MPP and NNGP, the AAGP still outperforms them. In the second and third scenarios, AAGP is able to give more robust prediction results than MPP and NNGP, since the true covariance function that MPP and NNGP are approximating is not the true covariance function used in the simulated field. AAGP seems to give more accurate results under a misspecified covariance function model, and this is especially important when statistical models are applied to analyze complicated spatio-temporal processes with unknown covariance structures. We also noticed that AAGP does not provide an exact approximation for the true covariance function. Although the discrepancy between the true covariance function and the covariance function resulting from AAGP cannot be eliminated, the second component in AAGP can greatly improve the prediction results compared to its computational-complexity-reduction method alone. In addition, the inference procedure in AAGP provides a computational strategy to make fast Bayesian inference when a computational-complexity-reduction method and a separable covariance function
model are combined.

The computing time for AAGP is roughly twice the computing time for MPP when they use the same number of knots. AAGP is much faster than the Full GP model and the NNGP model. It is worth mentioning that constructing a covariance matrix is slow and unavoidable in all these models. The computing time in the Full GP model shown in Table 3 is much larger than either the computing time in Table 1 or the computing time in Table 2 since constructing two correlation matrices $R_0, R_2$ takes more time than just constructing one correlation matrix $R_1$ or $R_2$. In addition, constructing the correlation matrices $R_1, R_*$ takes about 30% of the time in one MCMC iteration in MPP, since these matrices need to be evaluated five times in one MCMC iteration. This unavoidable time can potentially make MPP as well as AAGP slow. For the NNGP model, as noted in [Finley et al. (2017)] that sequential updating MCMC algorithm can be very slow, but implementation of more efficient MCMC algorithms for NNGP is beyond the scope of this paper.

4.2 Simulation example with a deterministic function

Section 4.2 shows that there is inferential and computational benefit to use AAGP when the underlying correlation structure is known. What follows is to investigate the performance of AAGP when we fit an unknown spatio-temporal function/surface. To see whether there is any inferential and computational benefit to use AAGP, we consider the function $f(s, u) = 10s \exp(-s^2 + u^2)) \sin(u/4)$, where $x = (s, u) \in \mathcal{X} = [-4, 4] \times [-4, 4]$. It is very challenging to fit this deterministic function and to model its spatio-temporal interaction. We first simulated data with the function $f(\cdot, \cdot)$ evaluated at 50-by-50 grids in the domain $\mathcal{X}$. Then we randomly held out 10% for prediction. The remaining 90% of data are used for parameter estimation. We compared the following methods:

1. Full GP with Gneiting’s nonseparable correlation function; (Gneiting)
2. Full GP with Gneiting’s nonseparable correlation function and separable correlation function
with exponential correlation functions for the first and second dimension; (Gneiting + SE)

(3) MPP with Gneiting’s nonseparable correlation function based on 200 and 400 knots;
(4) NNGP with Gneiting’s nonseparable correlation function based on 15 nearest neighbors;
(5) AAGP with Gneiting’s nonseparable correlation function and separable correlation function
with exponential correlation functions for the first and second dimension.

The knots in MPP and AAGP are selected based on nested hyper-latin cube design. 200 knots
in MPP and AGP are used, and additional 200 knots are added to MPP such that the computa-
tional time in MPP matches AAGP approximately. NNGP is implemented based on 15 nearest
neighbors. In what follows we focus on Gneiting’s nonseparable correlation function when fitting
each model for this deterministic function, although other nonseparable correlation can be used
as well. However, when data size is large, various computational-complexity-reduction methods
such as MPP, FSA, NNGP, and MRA have been proposed to approximate this single target corre-
lation function. So, the Full GP model with a single correlation function represents the best result
that these computational-complexity-reduction will achieve. In contrast, AAGP not just incorpo-
rates the correlation function inherited from these computational-complexity-reduction methods
but also includes a separable correlation function to capture what has been unexplained due to
different kinds of approximations. The Full GP model with two correlation functions is served as
the true model of what AAGP will approximate, so this will give the best results that one can hope
with AAGP.

For each model, the MCMC algorithms are run 25,000 iterations with a burn-in time of 15,000
iterations indicating independence based on convergence diagnostics. The predictive results and
computing time are reported for each method in Table 4. We can see that Full GP with a single
and two correlation functions gives the smaller MSPE than all the other methods, but Full GP with
the combination of Gneiting’s nonseparable correlation function and separable correlation function
gives smallest MSPE than the one only with Gneiting’s nonseparable correlation function. This
shows that adding a separable correlation function model to a nonseparable one will give improved inferential result. In practice, it is computationally prohibitive to implement these two models for large spatio-temporal datasets, so various computational-complexity-reduction methods will be used. In this example, MPP and NNGP are implemented, which give larger MSPE than AAGP. This suggests that by incorporating the additional separable correlation function, AAGP can give better results than any computational-complexity-reduction method itself.

### 4.3 Analysis of Eastern US ozone data

Ground-level ozone (O$_3$) is one of six common air pollutants identified in the Clean Air Act, and these air pollutants are called “criteria pollutants” by the U.S. Environmental Protection Agency (EPA). To protect human health and the environment, EPA publishes the National Ambient Air Quality Standards (NAAQS) for ozone, which specifies the maximum allowed measurement for ozone to be present in the outdoor air. The NAAQS for ozone is calculated based on the following steps: 1) the maximum 8-hour average is calculated for each day; 2) then the fourth-highest value is computed for these daily maximum 8-hour averages; 3) finally, the NAAQS for ozone is defined as the average of these fourth-highest values for any consecutive three-year period. The proposed method is illustrated with daily maximum 8-hour average data at a network of monitoring sites in the Eastern U.S. from April through October in the year 1995 to 1999. This data has been widely used in environmental statistics (see, for example, Fuentes, 2003; Gilleland and Nychka, 2005; Zhang et al., 2015b), and can be obtained on the website at [https://www.image.ucar.edu/Data/Ozmax](https://www.image.ucar.edu/Data/Ozmax). Following the pre-processing steps in Gilleland and Nychka (2005), the daily maximum 8-hour ozone average at station $s$ and day $u$, denoted by $O(s, u)$, is assumed to have the following structure $O(s, u) = \mu(s, u) + \tilde{O}(s, u)$, where $\mu(s, u) = a(s) + \sum_{j=1}^{3} \{ b_j \cos(2\pi ju/184) + c_j \sin(2\pi ju/184) \}$, which models the seasonal effect. The coefficients in the seasonal effect $\mu(s, u)$ are estimated through ordinary least square method.
The spatial-varying standard deviation $k(\cdot)$ is estimated with the empirical standard deviation based on residuals after removing the seasonal effect. The residual $r(s, u) \equiv O(s, u) - \hat{\mu}(s, u)$ scaled by its estimated standard deviation $\hat{k}(\cdot)$ at each station is referred to as standardized ozone at station $s$ and time $u$ hereafter. The spatio-temporal model (2.1) with zero mean is assumed for the standardized ozone, based upon which statistical inference is carried out.

The datasets used in this paper are obtained at 513 monitoring sites during 92 days from June to August in 1997, where $1.37\% = 645/(513 \times 92) \times 100\%$ of data are missing, and only 46551 data points are observed. In what follows, statistical analysis is carried out on the standardized data using the previously mentioned pre-processing steps. To analyze these data, a cross-validation is first carried out on observed 46551 data points on June to August in 1997, where 90% randomly selected data points are used for parameter estimation, and the remaining 10% data points are used to assess predictive performance. In the cross-validation, three covariance function models are compared: MPP and NNGP with Gneiting’s space-time covariance function, and the AAGP with Gneiting’s space-time covariance function and exponential covariance functions in the separable covariance function. Based on exploratory analysis, the prior distributions are specified as $\sigma_1^2 \sim IG(2, 0.01), \sigma_2^2 \sim IG(2, 0.01), \tau^2 \sim IG(2, 0.01), a \sim U(0, 60), c \sim U(0, 2000), \beta \sim U(0, 1)$ in Gneiting’s space-time correlation function and $\phi_s \sim U(0, 2000), \phi_u \sim U(0, 60)$ in separable covariance functions in space and time. In MPP and AAGP, 490 knots are selected in the spatio-temporal domain via hyper Latin design. Then we further increase the number of knots up to 1200 in MPP to investigate whether AAGP with just 490 still outperform MPP. The distance in space is calculated based on chordal distance, and the distance in time is calculated based on Euclidean distance. The NNGP model is implemented with 15 nearest neighbors.

For the cross-validation, the results in Table 5 show that AAGP gives better predictive performance than MPP even though more knots are added. In addition, AAGP also gives better prediction results than NNGP. The variance for Gneiting’s nonseparable covariance function in MPP is much
larger than the variance estimated in AAGP, but the overall variance is estimated consistently based on MPP and AAGP. There are two spatial range parameters $c, \phi_s$ and two temporal range parameters $a, \phi_t$ in AAGP. The meaning of these parameters should not be confused with corresponding parameters in the covariance function model with just only one of the components. Although the time in AAGP is longer than MPP with 490 knots (roughly a factor of 2.3), the improved predictive accuracy is noticeable. Compared with NNGP, the computational time for AAGP is smaller. But our sequential implementation of NNGP is not computationally efficient (e.g., Finley et al., 2017) although these two methods are both implemented in the same software platform. We would also like to point out that the computational-complexity-reduction method in AAGP is derived from MPP, but other methods such as NNGP or MRA can also be used in AAGP to further achieve inferential benefit.

Predictions are also carried out over space for different days based on all observed data. Figure 1 visualizes the predictions on three consecutive days based on all available data, which clearly shows that AAGP is able to capture the spatio-temporal dependence structures in the data.

5 Discussion

In this paper, we have proposed a new additive Gaussian process approximation scheme for large spatio-temporal datasets based on the combination of a computational-complexity-reduction covariance function and a separable covariance function. The proposed approximation method AAGP provides a flexible way to characterize spatio-temporal dependence structures and is able to give better prediction results. We also proposed a fully conditional Markov chain Monte Carlo algorithm based on the hierarchical representation of the model. This proposed Bayesian inference framework allows efficient computations for large spatio-temporal data, and avoids expensive calculation of the marginal likelihood. In addition, the new approximation method is studied under different simulation scenarios, and we show that it can give good prediction results under vari-
Figure 1. Standardized ozone data and predictions for $Y(\cdot)$ based on 30 by 40 locations on three consecutive days (unit: parts per million). The panels (a), (d), (g) show the standardized ozone on June 14, 15, 16 in 1997. The panels (b), (e), (h) show the posterior mean on June 14, 15, 16 in 1997. The panels (c), (f), (i) show the corresponding posterior standard errors.

ous covariance structures as well as a deterministic function. We also applied this approximation method to analyze Eastern U.S. ozone data.

MPP, chosen as a computational-complexity-reduction method in AAGP, can be considered as one instance of the general Vecchia approximations as shown in Katzfuss and Guinness (2017), thus the AAGP method can be extended to embed other types of the general Vecchia approximations including FSA, NNGP, MRA and sparse general Vecchia (SGV). This will lead to a more
general approximation method. The corresponding Bayesian inference can also be extended to incorporate these methods to allow fast computations for large spatio-temporal data. A more rigorous algorithmic development as well as comparison has been left for future work. Moreover, the efficiency of the proposed Bayesian inference procedure can be improved using partially collapsed Gibbs samplers (van Dyk and Park [2008], van Dyk and Jiao [2015]).

Finally, the AAGP method relies on a computational-complexity-reduction covariance function and a separable covariance function, which typically allows fast computation for large spatio-temporal datasets. When the number of data points in space (or time) is large and the number of data points in time (or space) is small, one can also incorporate a purely spatial or temporal computational-complexity-reduction method. The proposed fully conditional Markov chain Monte Carlo algorithm can also be applied efficiently.

References

Ba, S. and Joseph, V. R. (2012). Composite Gaussian process models for emulating expensive functions. *The Annals of Applied Statistics*, 6(4):1838–1860.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*. CRC Press.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848.

Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons, New York, revised edition.

Cressie, N. and Huang, H.-C. (1999). Classes of Nonseparable, Spatio-Temporal Stationary Covariance Functions. *Journal of the American Statistical Association*, 94(448):1330–1340.
Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. John Wiley & Sons, New York.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016a). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.

Datta, A., Banerjee, S., Finley, A. O., Hamm, N. A. S., and Schaap, M. (2016b). Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *The Annals of Applied Statistics*, 10(3):1286–1316.

De Iaco, S., Myers, D. E., and Posa, D. (2002). Nonseparable space-time covariance models: Some parametric families. *Mathematical Geology*, 34(1):23–42.

Finley, A. O., Datta, A., Cook, B. C., Morton, D. C., Andersen, H. E., and Banerjee, S. (2017). Efficient algorithms for Bayesian nearest neighbor Gaussian processes. *arXiv:1702.00434*.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics and Data Analysis*, 53(8):2873–2884.

Fonseca, T. C. O. and Steel, M. F. J. (2011). A general class of nonseparable space-time covariance models. *Environmetrics*, 22(2):224–242.

Fuentes, M. (2003). Statistical assessment of geographic areas of compliance with air quality standards. *Journal of Geophysical Research*, 108(D):9002.
Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409.

Genton, M. G. (2007). Separable approximations of space-time covariance matrices. *Environmetrics*, 18(7):681–695.

Gilleland, E. and Nychka, D. (2005). Statistical models for monitoring and regulating ground-level ozone. *Environmetrics*, 16(5):535–546.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97(458):590–600.

Gramacy, R. B. and Apley, D. W. (2015). Local Gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, 24(2):561–578.

Guhaniyogi, R., Finley, A. O., Banerjee, S., and Gelfand, A. E. (2011). Adaptive gaussian predictive process models for large spatial datasets. *Environmetrics*, 22(8):997–1007.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Henderson, H. V. and Searle, S. R. (1981). On deriving the inverse of a sum of matrices. *SIAM Review*, 23(1):53–60.

Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112(517):201–214.

Katzfuss, M. and Guinness, J. (2017). A general framework for Vecchia approximations of Gaussian processes. *arXiv:1708.06302*.

Kaufman, C. G. and Shaby, B. A. (2013). The role of the range parameter for estimation and prediction in geostatistics. *Biometrika*, 100(2):473–484.
Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Ma, P. and Kang, E. L. (2017). Fused Gaussian process for very large spatial data. *arXiv:1702.08797*.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015). A multiresolution Gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, 24(2):579–599.

Rougier, J. (2008). Efficient emulators for multivariate deterministic functions. *Journal of Computational and Graphical Statistics*, 17(4):827–843.

Sang, H. and Huang, J. Z. (2012). A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):111–132.

Stein, M. L. (2005). Space–time covariance functions. *Journal of the American Statistical Association*.

Stein, M. L., Chi, Z., and Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):275–296.

van Dyk, D. A. and Jiao, X. (2015). Metropolis-Hastings within partially collapsed Gibbs samplers. *Journal of Computational and Graphical Statistics*, 24(2):301–327.

van Dyk, D. A. and Park, T. (2008). Partially collapsed Gibbs samplers. *Journal of the American Statistical Association*, 103(482):790–796.
Zhang, B., Konomi, B. A., Sang, H., Karagiannis, G., and Lin, G. (2015a). Full scale multi-output Gaussian process emulator with nonseparable auto-covariance functions. *Journal of Computational Physics*, 300(C):623–642.

Zhang, B., Sang, H., and Huang, J. Z. (2015b). Full-scale approximations of spatio-temporal covariance models for large datasets. *Statistica Sinica*, 25(1):99–114.

Zhang, B., Sang, H., and Huang, J. Z. (2018). Smoothed full-scale approximation of Gaussian process models for computation of large datasets. *Statistica Sinica*, Accepted.
Table 1. Simulation results under a nonseparable space-time correlation function

| Parameters | True value | Full GP  | MPP $m = 250$ | NNGP $m = 704$ | AAGP $m = 15$ | AAGP $m = 250$ |
|------------|------------|----------|----------------|----------------|----------------|----------------|
| $b_1$      | 1          | 0.989(0.980, 0.998) | 0.990(0.981, 1.000) | 0.993(0.968, 1.019) | 1.033(0.717, 1.303) | 0.981(0.959, 1.007) |
| $b_2$      | 0.5        | 0.503(0.473, 0.533) | 0.503(0.491, 0.516) | 0.482(0.445, 0.517) | 0.495(0.382, 0.601) | 0.507(0.468, 0.534) |
| $\sigma_1^2$ | 1          | 0.978(0.869, 1.096) | 1.419(1.236, 1.645) | 1.253(1.070, 1.474) | 1.062(0.905, 1.241) | 0.869(0.787, 1.038) |
| $\beta$    | 0.8        | 0.841(0.425, 0.994) | 0.935(0.710, 0.999) | 0.913(0.549, 0.995) | 0.734(0.169, 0.992) | 0.928(0.631, 0.997) |
| $\sigma_2^2$ | 1          | 0.991(0.754, 1.255) | 2.070(1.613, 2.659) | 2.301(1.728, 3.073) | 0.738(0.541, 0.957) | 2.805(2.054, 3.872) |
| $\sigma_3^2$ | 5          | 4.768(4.022, 5.710) | 5.465(4.503, 6.560) | 7.700(6.107, 9.737) | 5.302(4.327, 6.573) | 5.542(4.685, 6.750) |
| $\phi_s$   | 0.2        | 0.182(0.146, 0.221) | 0.130(0.049, 0.208) | 0.216(0.154, 0.288) | 0.189(0.157, 0.219) | 0.060(0.021, 0.130) |
| $\phi_t$   | 0.5        | 0.503(0.473, 0.533) | 0.503(0.491, 0.516) | 0.482(0.445, 0.517) | 0.495(0.382, 0.601) | 0.507(0.468, 0.534) |
| $\tau^2$   | 0.2        | 0.182(0.146, 0.221) | 0.130(0.049, 0.208) | 0.216(0.154, 0.288) | 0.189(0.157, 0.219) | 0.060(0.021, 0.130) |

Table 2. Simulation results under separable correlation function

| Parameters | True value | Full GP  | MPP $m = 250$ | NNGP $m = 704$ | AAGP $m = 15$ | AAGP $m = 250$ |
|------------|------------|----------|----------------|----------------|----------------|----------------|
| $b_1$      | 1          | 0.989(0.975, 1.004) | 1.003(0.973, 1.033) | 0.995(0.962, 1.027) | 0.877(0.607, 1.076) | 0.989(0.974, 1.003) |
| $b_2$      | 0.5        | 0.503(0.483, 0.522) | 0.498(0.455, 0.542) | 0.493(0.447, 0.539) | 0.490(0.445, 0.598) | 0.503(0.484, 0.522) |
| $\sigma_1^2$ | 1          | 0.083(0.045, 0.163) | 0.052(0.027, 0.098) | 0.705(0.635, 0.807) | 0.030(0.012, 0.060) | 0.837(0.372, 0.993) |
| $\sigma_2^2$ | 0.737(0.104, 0.987) | 0.718(0.083, 0.989) | 0.498(0.026, 0.974) | 0.837(0.372, 0.993) | 0.718(0.083, 0.989) | 0.498(0.026, 0.974) |
| $\sigma_3^2$ | 17.81(9.540, 20.000) | 19.08(11.37, 20.000) | 7.086(6.267, 8.645) | 0.198(0.023, 0.835) | 7.086(6.267, 8.645) | 0.198(0.023, 0.835) |
| $\phi_s$   | 5          | 4.968(4.646, 5.213) | 4.968(4.646, 5.213) | 4.968(4.646, 5.213) | 4.968(4.646, 5.213) | 4.968(4.646, 5.213) |
| $\phi_t$   | 1          | 0.990(0.918, 1.067) | 0.990(0.918, 1.067) | 0.990(0.918, 1.067) | 0.990(0.918, 1.067) | 0.990(0.918, 1.067) |
| $\tau^2$   | 0.2        | 0.188(0.179, 0.196) | 0.950(0.992, 1.037) | 1.125(1.072, 1.180) | 0.109(0.096, 0.124) | 0.187(0.178, 0.196) |

MSPE 0.02 0.58 0.48 0.69 0.69
ALCI 0.58 0.58 0.58 0.58 0.58
Time (h) 40.9 40.9 40.9 40.9 40.9
### Table 3. Simulation results under a combination of nonseparable and separable correlation functions

| Third Scenario | Parameters |        | Full GP |        | MPP |        | NNGP |        | AAGP |
|----------------|------------|--------|---------|--------|-----|--------|------|--------|------|
|               | True value |        |         |        |     |        | 15   |        |      |
| $b_1$         | 1          | 0.995(0.944, 0.999) | 0.969(0.932, 1.007) | 0.965(0.930, 1.001) | 0.822(0.390, 1.209) | 0.969(0.949, 0.997) |
| $b_2$         | 0.5        | 0.502(0.472, 0.532) | 0.499(0.446, 0.552) | 0.463(0.412, 0.511) | 0.490(0.345, 0.612) | 0.509(0.475, 0.540) |
| $\sigma_1^2$  | 1          | 0.993(0.863, 1.204) | 1.615(1.319, 1.957) | 1.915(1.611, 2.240) | 2.290(2.245, 2.300) | 1.356(1.181, 1.463) |
| $\beta$       | 0.8        | 0.847(0.396, 0.994) | 0.897(0.539, 0.996) | 0.850(0.447, 0.994) | 0.172(0.005, 0.562) | 0.848(0.332, 0.993) |
| $a$           | 1          | 0.961(0.705, 1.316) | 2.379(1.324, 3.677) | 2.172(1.454, 3.223) | 3.220(2.714, 4.090) | 1.737(1.409, 2.248) |
| $c$           | 5          | 4.758(3.857, 6.226) | 4.973(3.660, 6.639) | 3.850(2.885, 5.081) | 7.781(6.853, 8.964) | 8.253(6.711, 10.43) |
| $\sigma_2^2$  | 1          | 0.916(0.730, 1.172) | 0.970(0.806, 1.154) | 0.949(0.874, 1.035) | 0.969(0.949, 0.997) | 0.969(0.949, 0.997) |
| $\phi_s$      | 5          | 4.889(4.529, 5.213) | 5.047(4.878, 5.150) | 5.047(4.878, 5.150) | 5.047(4.878, 5.150) | 5.047(4.878, 5.150) |
| $\phi_u$      | 1          | 1.003(0.926, 1.071) | 1.037(0.989, 1.095) | 1.037(0.989, 1.095) | 1.037(0.989, 1.095) | 1.037(0.989, 1.095) |
| $\tau^2$      | 0.2        | 0.175(0.141, 0.211) | 0.583(0.236, 0.795) | 0.368(0.078, 0.567) | 0.504(0.454, 0.559) | 0.504(0.454, 0.559) |
| MSPE          |            | 0.28    | 1.40    | 1.30   | 0.83  | 0.48   |
| ALCI          |            | 2.10    | 3.85    | 4.03   | 3.52  | 3.40   |
| Time (h)      |            | 106     | 1.44    | 4.00   | 4.34  | 3.47   |

### Table 4. Simulation results with a deterministic function.

| Model |     | Full GP |     | MPP |     | NNGP |     | AAGP |
|-------|-----|---------|-----|-----|-----|------|-----|------|
|       |     | Gneiting | Gneiting + SE | m = 200 | m = 400 | 15 | m = 250 |
| MSPE  | $8.2 \times 10^{-7}$ | $7.2 \times 10^{-7}$ | $8.4 \times 10^{-4}$ | $6.1 \times 10^{-5}$ | $1.5 \times 10^{-5}$ | $5.2 \times 10^{-6}$ |
| ALCI  | 0.03 | 0.03 | 0.13 | 0.08 | 0.05 | 0.06 |
| Time (h) | 14.5 | 22.3 | 1.15 | 3.07 | 2.30 | 2.10 |

### Table 5. Cross validation results for standardized ozone from June to August, 1997. Gneiting’s nonseparable correlation function is used in all these models. Exponential correlation functions are used for the separable covariance function in AAGP.

| Parameters | MPP | NNGP | AAGP |
|------------|-----|------|------|
| $m = 490$  |     |      |      |
| $m = 1200$ |     |      |      |
| $\sigma_1^2$ | 0.853(0.796, 0.914) | 0.949(0.874, 1.035) | 0.932(0.897, 0.983) |
| $\beta$     | 0.973(0.996, 1.000) | 0.983(0.907, 0.999) | 0.503(0.025, 0.9798) |
| $\alpha$ (day) | 1.674(1.557, 1.782) | 1.084(0.992, 1.170) | 1.438(1.285, 1.597) |
| $c$ (km)    | 1508(1406, 1625) | 1311(1202, 1444) | 388.5(377.1, 412.6) |
| $\sigma_2^2$ | 1.084(0.992, 1.170) | 1.438(1.285, 1.597) | 5.196(3.932, 7.112) |
| $\phi_s$ (km) | 480.1(478.4, 481.8) | 558.2(556.5, 559.9) | 636.3(634.6, 637.8) |
| $\phi_u$ (day) | 1.85(1.78, 1.93) | 1.85(1.78, 1.93) | 1.85(1.78, 1.93) |
| $\tau^2$    | 0.194(0.187, 0.202) | 0.194(0.187, 0.202) | 0.194(0.187, 0.202) |
| MSPE        | 0.41 | 0.16 | 0.04 |
| ALCI        | 1.88 | 1.44 | 1.05 |
| Time (h)    | 41.8 | 91.8 | 87.7 |