

# Circulant embedding of approximate covariances for inference from Gaussian data on large lattices

Joseph Guinness and Montserrat Fuentes

## Abstract

Recently proposed computationally efficient Markov chain Monte Carlo and Monte Carlo Expectation-Maximization (EM) methods for estimating covariance parameters from lattice data rely on successive imputations of values on an embedding lattice that is at least two times larger in each dimension. These methods can be considered exact in some sense, but we demonstrate that using such a large number of imputed values leads to slowly converging Markov chains and EM algorithms. We propose instead the use of a discrete spectral approximation to allow for the implementation of these methods on smaller embedding lattices. While our methods are approximate, our examples indicate that the error introduced by this approximation is small compared to the Monte Carlo errors present in long Markov chains or many iterations of Monte Carlo EM algorithms. Our results are demonstrated in simulation studies, as well as in numerical studies that explore both increasing domain and fixed domain asymptotics. We compare the exact methods to our approximate methods on a large satellite dataset, and show that the approximate methods are also faster to compute, especially when the aliased spectral density is modeled directly.

## 1 Introduction

The Gaussian process model plays a central role in the analysis of spatially and spatial-temporally correlated data. It is used directly for modeling data that can be assumed to be Gaussian and often used indirectly as a stage in a hierarchical process model when the data are not assumed to be Gaussian (Banerjee et al., 2014). Consider a stochastic process  $Z(\mathbf{x}) \in \mathbb{R}$ ,  $\mathbf{x} \in \mathbb{R}^d$ . The defining property of a Gaussian process is that for any  $n \in \mathbb{N}$  and  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ , the vector  $\mathbf{Z} = (Z(\mathbf{x}_1), \dots, Z(\mathbf{x}_n))'$  has a multivariate normal distribution. A Gaussian process is characterized by its mean at every location  $E(Z(\mathbf{x}))$  and the covariance between its observations at any two locations  $\text{Cov}(Z(\mathbf{x}), Z(\mathbf{y}))$ , and it is common to assume that both the means and the covariances are specified by parametric functions. We write  $\boldsymbol{\mu}_\beta = E(\mathbf{Z})$  and  $K_\theta = E((\mathbf{Z} - \boldsymbol{\mu}_\beta)(\mathbf{Z} - \boldsymbol{\mu}_\beta)')$  to signify a mean vector with parameter  $\beta$  and covariance

matrix with parameter  $\theta$  for observations at a specific set of locations. Statistical computation with Gaussian process models requires algebraic manipulations involving  $\boldsymbol{\mu}_\beta$  and  $K_\theta$ . The Cholesky factorization is commonly used in computing the Gaussian log-likelihood function when  $K_\theta$  has no exploitable structure. The Cholesky factorization requires  $O(n^3)$  floating point operations (flops) and  $O(n^2)$  memory, so its computational burden begins to overwhelm modern standard computational facilities when  $n$  is greater than 10,000. The addition or subtraction of the mean vector requires only  $O(n)$  flops and memory, so we focus our attention here on computations involving covariances and assume throughout that the mean vector is zero.

Stroud et al. (2014) recently proposed methods that leverage the computational power of circulant embedding techniques (Wood and Chan, 1994; Dietrich and Newsam, 1997; Stein, 2002; Gneiting et al., 2006) to make inferences about parameters in stationary Gaussian random field models when data are observed on a possibly incomplete lattice. The methods rely on MCMC and Monte Carlo EM algorithms that successively impute data to a larger periodic lattice, on which the Gaussian loglikelihood can be computed efficiently with Fast Fourier Transform (FFT) algorithms. While the methods in Stroud et al. (2014) are exact, in that the model for the data assumed by the algorithms is equal to the target covariance model, the methods require an embedding lattice that is much larger than the observation lattice, resulting in at least 8 times as many imputed data points as actual observations. In our paper, we note that when approximations of the covariance are employed, the number of imputed data points can be reduced substantially. We demonstrate that this reduction in imputed values has a dramatic impact on the computational efficiency of the iterative algorithms for estimating covariance parameters, and while the inferences are technically approximate, we can actually improve parameter estimation over exact methods in the context of a fixed budget of computational time. We briefly review the discrete spectral approximation in Section 2, and we introduce efficient methods for computing it when the unaliased spectral density or the covariance function is specified rather than the aliased spectral density. In Section 3, we provide numerical experiments studying the accuracy of the discrete spectral approximation for the purpose of estimating covariance parameters via likelihood methods, considering varying strengths of spatial correlation as well as fixed- and increasing-domain scenarios. We provide somewhat surprising examples in which the asymptotic biases of our parameter estimates are small even when some of the covariances at large lags are not well approximated. The simulation studies in Section 4 demonstrate that MCMC and Monte Carlo EM algorithms converge more quickly, and with no noticeable loss in accuracy when the discrete spectral approximation is used, as long as the lattice is expanded by at least a factor of 5/4 in each dimension. The methods are further compared in Section 5 on oceanic satellite data, where we observe that the approximate methods are computed faster than the exact methods, especially when the aliased spectral density is modeled directly. We conclude with a discussion in Section 6.

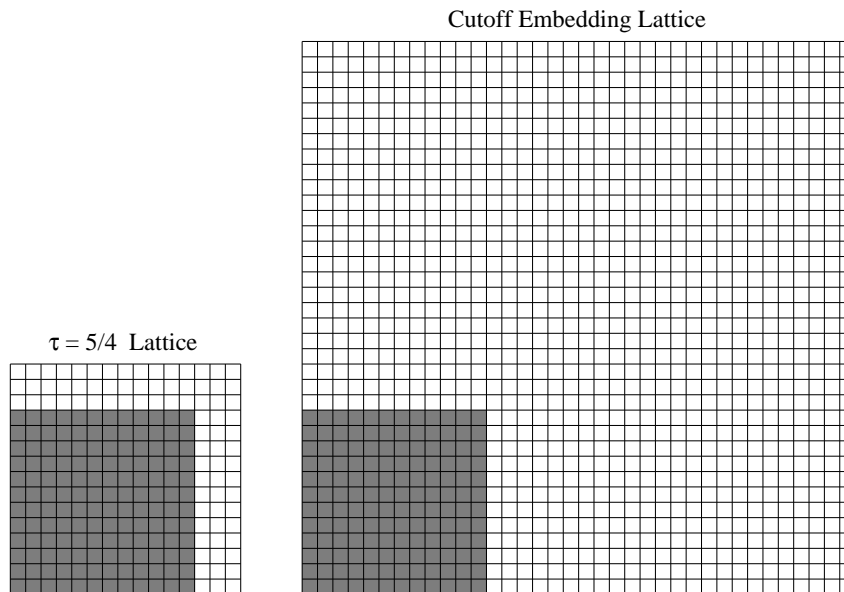


Figure 1: Illustration of the relative sizes of the embedding lattices ( $\delta\mathbb{J}_{\mathbf{m}}$ , gray+white) and the observation lattice ( $\delta\mathbb{J}_{\mathbf{n}}$ , gray) used in our approximate procedures with  $\tau = 5/4$  and in cutoff embedding, which uses  $\tau = 3$  in Stroud et al. (2014). Here, the observation lattice has size  $\mathbf{n} = (12, 12)$ .

## 2 Discrete Spectral Approximation

We write  $\delta\mathbb{Z}^d$  to denote a regular  $d$ -dimensional lattice with spacing  $\delta > 0$ , meaning that  $\mathbf{x} = (x_1, \dots, x_d) \in \delta\mathbb{Z}^d$  if  $x_j/\delta \in \mathbb{Z}$  for every  $j$ . Let  $\mathbf{n} = (n_1, \dots, n_d)$  be a vector of natural numbers, and define  $\delta\mathbb{J}_{\mathbf{n}}$  to be a rectangular subset of  $\delta\mathbb{Z}^d$  with  $n_j$  points along dimension  $j$ . We refer to  $\delta\mathbb{J}_{\mathbf{n}}$  as the observation lattice. We also define a larger embedding lattice  $\delta\mathbb{J}_{\mathbf{m}}$ , so that each component of  $\mathbf{m}$  is larger than the corresponding component of  $\mathbf{n}$ , and we define the ratio  $\tau_j = m_j/n_j \geq 1$ . Exact embedding methods typically require  $\tau_j \geq 2$ , and the particular method described in Stroud et al. (2014, Section 4.1)—a variant of the cutoff embedding methods described in Stein (2002) and Gneiting et al. (2006)—requires  $\tau_j \geq 2\sqrt{d}$ . In this section, we describe the discrete spectral approximation, which allows us to construct covariance functions that are always positive definite and periodic on  $\delta\mathbb{J}_{\mathbf{m}}$  with  $\tau_j < 2$ . In Figure 1, we illustrate the sizes of the embedding lattices used in the discrete spectral approximation with  $\tau = 5/4$  and in typical applications with cutoff embedding.

We assume throughout that the covariance function  $K(\mathbf{h})$  has a spectral representation

$$K(\mathbf{h}) = \int_{\mathbb{R}^d} \exp(i\boldsymbol{\omega}'\mathbf{h})f(\boldsymbol{\omega})d\boldsymbol{\omega}, \quad (1)$$

in terms of a continuous spectral density  $f(\boldsymbol{\omega})$ . Existence of the spectral density is

a usual assumption in circulant embedding literature; for example, Wood and Chan (1994), Dietrich and Newsam (1997), and Gneiting et al. (2006) all make this assumption. Continuity of  $f$  is a slightly stronger assumption, which we use in Lemma C.1 of the Supplementary Material. to obtain theoretical results about our methods for computing the discrete spectral approximation when the covariance function is specified. When data are observed on a regular lattice  $\delta\mathbb{Z}^d$ , we need to consider only covariances  $K(\mathbf{h})$  with  $\mathbf{h} \in \delta\mathbb{Z}^d$ , and thus the covariance function can be expressed as

$$K(\mathbf{h}) = \sum_{\mathbf{j} \in \mathbb{Z}^d} \int_{[0, 2\pi/\delta]^d} f(\boldsymbol{\omega} + 2\pi\mathbf{j}/\delta) \exp(i(\boldsymbol{\omega} + 2\pi\mathbf{j}/\delta)'\mathbf{h}) d\boldsymbol{\omega} = \int_{[0, 2\pi/\delta]^d} f_\delta(\boldsymbol{\omega}) \exp(i\boldsymbol{\omega}'\mathbf{h}) d\boldsymbol{\omega}, \quad (2)$$

with the second equality following by exchanging summation and integration and using the fact that  $\exp(i\boldsymbol{\omega}'\mathbf{h})$  is aliased with  $\exp(i(\boldsymbol{\omega} + 2\pi\mathbf{j}/\delta)'\mathbf{h})$  on  $\mathbf{h} \in \delta\mathbb{Z}^d$ . We call  $f_\delta(\boldsymbol{\omega}) = \sum_{\mathbf{j} \in \mathbb{Z}^d} f(\boldsymbol{\omega} + 2\pi\mathbf{j}/\delta)$  the aliased spectral density. The discrete spectral approximation is a discretization of the integral in (2),

$$R_{\mathbf{m}}(\mathbf{h}) = \frac{(2\pi)^d}{m} \sum_{\mathbf{j} \in \mathbb{J}_{\mathbf{m}}} f_\delta(\boldsymbol{\omega}_{\mathbf{j}}) \exp(i\boldsymbol{\omega}'_{\mathbf{j}}\mathbf{h}), \quad (3)$$

where  $\boldsymbol{\omega}_{\mathbf{j}} = (2\pi j_1/(\delta m_1), \dots, 2\pi j_d/(\delta m_d))$  are the  $d$ -dimensional Fourier frequencies on a grid of size  $\mathbf{m}$ . By varying the size of  $\mathbf{m}$ —or equivalently,  $\tau$ —in  $R_{\mathbf{m}}(\cdot)$ , the practitioner has the ability to control its accuracy. The use of discrete spectral approximations for simulating Gaussian processes has a long history; (Cressie, 1993, pg. 204-205) gives a brief historical account. The approximation in (3) is powerful since  $R_{\mathbf{m}}(\cdot)$  is automatically periodic with period  $m_j$  in each dimension, efficient to compute with FFT algorithms when  $f_\delta(\boldsymbol{\omega}_{\mathbf{j}})$  is available, and guaranteed to be positive definite, whereas the variant of cutoff embedding employed in Stroud et al. (2014, Section 4.1) is not guaranteed to be positive definite.

The covariance function in (3) is periodic on the embedding lattice  $\delta\mathbb{J}_{\mathbf{m}}$ , so the resulting covariance matrix  $R_{\mathbf{m}}$  for the set of all values on  $\delta\mathbb{J}_{\mathbf{m}}$  (ordered lexicographically) has a nested block circulant structure (as defined in Appendix A of the Supplementary Material). Rue and Held (2005) and Lindgren et al. (2011) discuss methods for constructing Markov random field models that are periodic on a domain that is slightly larger than the observation domain. Here, we do not assume that the random field has a Markov or approximately Markov structure, only that it is stationary. Indeed, we consider examples with the exponential covariance function, which is not Markov in two dimensions. We propose to use the Bayesian and frequentist methods in Stroud et al. (2014). The difference here is that we replace  $K_\theta$  with  $R_{\mathbf{m}}$  (which depends on  $\theta$  as well). We show in Sections 4 and 5 that this replacement has dramatic impacts on the computational efficiency of the methods without sacrificing the accuracy of the parameter estimates in the examples we study. The tradeoff of our approach is its approximate nature, but the advantage is that since the approximations

can often be made very sharp with  $\tau < 2$ , the number of imputed values required for the Monte Carlo methods is not large compared to the number of observed values, allowing the observed values to drive the parameter updates. We demonstrate the accuracy of this approximation for small  $\tau$  in Section 3. The simulations in Section 4 show that the computational advantage of using fewer imputed values far outweighs the drawback of the approximate nature of the method, and in Section 5, we show that the overall computational effort is smallest when  $f_\delta(\cdot)$  is modeled directly with flexible elementary parametric functions.

## 2.1 Computation of $R_{\mathbf{m}}$

In this subsection we discuss three methods for computing and approximating the covariances  $R_{\mathbf{m}}(\mathbf{h})$ . We show in Appendix C of the Supplementary Material that

$$R_{\mathbf{m}}(\mathbf{h}) = \sum_{\mathbf{j} \in \mathbb{Z}^d} K(\mathbf{h} + \delta \mathbf{j} \circ \mathbf{m}) = \lim_{N \rightarrow \infty} \sum_{j_1 = -N}^{N-1} \cdots \sum_{j_d = -N}^{N-1} K((h_1 + \delta j_1 m_1, \dots, h_d + \delta j_d m_d)), \quad (4)$$

where  $\mathbf{j} \circ \mathbf{m} = (j_1 m_1, \dots, j_d m_d)$ , so when  $K(\mathbf{h})$  decays quickly with  $\|\mathbf{h}\|$ , where  $\|\cdot\|$  denotes Euclidean distance, as in the exponential covariance function,  $R_{\mathbf{m}}(\mathbf{h})$  is well-approximated by a truncation of (4). This truncation does not guarantee positive definiteness, but commonly used covariance functions such as the powered exponential and the Matérn covariance functions (Matérn, 1960) decay quickly with  $\|\mathbf{h}\|$ , so the resulting covariance matrices rarely fail to be positive definite with  $N = 2$  or  $3$ , especially when the components of  $\mathbf{n}$  are large. The matrices in the examples we study in this paper never failed to be positive definite with  $N = 3$ , but if a matrix were not positive definite, we could simply increase  $N$  until the matrix became positive definite.

If the spectral density  $f(\boldsymbol{\omega})$  is available in closed-form, and it decays quickly with  $\|\boldsymbol{\omega}\|$ , one may approximate  $f_\delta(\boldsymbol{\omega}_{\mathbf{j}})$  by truncating  $\sum_{\mathbf{k} \in \mathbb{Z}^d} f(\boldsymbol{\omega}_{\mathbf{j}} + 2\pi \mathbf{k}/\delta)$ . This approximation is guaranteed to generate a positive definite function. Most of the commonly-used covariance models in spatial statistics possess the property that either the spectral density or the covariance function decays quickly. The Matérn covariance function, for example, decays faster than any polynomial, as do the spectral densities for the Gaussian and Cauchy covariance functions.

In Section 5 we show that our methods are particularly efficient when the aliased spectral density  $f_\delta(\cdot)$  is modeled directly in closed form. Then the array of eigenvalues  $f_\delta(\boldsymbol{\omega}_{\mathbf{j}})$ ,  $\mathbf{j} \in \mathbb{J}_{\mathbf{m}}$ , can be formed directly without any truncations, and  $R_{\mathbf{m}}$  is thus guaranteed to be positive definite. We recommend this modeling approach due to its computational advantages, and we discuss in Section 5 a parametric model for  $f_\delta(\cdot)$  that mimics the flexibility of the Matérn covariance function.

### 3 Numerical Studies

The accuracy of the discrete spectral approximation depends on an expansion factor  $\tau$ . The numerical studies in this section concern the resulting approximate covariance matrix for the observations  $R$  and the effect that the number of observations, the strength of spatial correlation, and the choice of  $\tau$  have on how well  $R$  approximates the target covariance matrix  $K$ . Since both the Bayesian and frequentist methods discussed in this paper are likelihood-based, we study the approximations with respect to Kullback-Leibler (KL) divergences. Let  $\mathbf{Z}$  be the set of all observations on  $\delta\mathbb{J}_n$ . Defining  $L_\tau(\theta; \mathbf{Z})$  to be the Gaussian loglikelihood function for  $\mathbf{Z}$  under covariance function  $R_m(\cdot)$  with  $\mathbf{m} = \tau\mathbf{n}$ , and defining  $L(\theta; \mathbf{Z})$  to be the Gaussian loglikelihood function for  $\mathbf{Z}$  under the target model, which has autocovariance function  $K_\theta(\cdot)$ , the KL divergence of our approximate model from the target model is  $E_0(L(\theta_0; \mathbf{Z}) - L_\tau(\theta; \mathbf{Z}))$ , where the expectation is taken with respect to the target model with parameter  $\theta_0$ . The  $\theta$  that maximizes  $L_\tau(\theta; \mathbf{Z})$  is consistent for  $\theta^\tau$ , the minimizer of the KL divergence, under replication of  $\mathbf{Z}$  (Varin et al., 2011). Since  $E_0(L(\theta_0; \mathbf{Z}))$  does not depend on  $\theta$ ,  $\theta^\tau$  is the minimizer of

$$E_0(-L_\tau(\theta; \mathbf{Z})) = \frac{1}{2} \log \det R + \frac{1}{2} \text{tr}(R^{-1}K),$$

where  $R$  implicitly depends on  $\theta$  and  $\tau$ , and  $K$  is formed using the covariance function  $K_{\theta_0}(\cdot)$ . In this section, we compute  $\theta^\tau$  for various choices of  $\tau$  in two different asymptotic scenarios and for various values of true parameter  $\theta_0$ . While we do not propose using maximizers of  $L_\tau(\theta; \mathbf{Z})$  as estimators in practice, the results of these computations are nonetheless useful for understanding how the quality of the approximate covariance functions depends on  $\tau$ , the size of the observation lattice, and the strength of spatial correlation.

The numerical studies use the isotropic exponential covariance function  $K_\lambda(\mathbf{h}) = \exp(-\|\mathbf{h}\|/\lambda)$  with range parameter  $\lambda$ . All of the studies assume a square two-dimensional lattice. The lattice in the first set of calculations has spacing  $\delta = (32\sqrt{2})^{-1}$  and  $\lambda_0 = 0.15$ , and we specify lattice sizes of  $n \in \{32^2, 48^2, 64^2, 80^2\}$  an increasing domain scenario. The covariance approximations are calculated with values of  $\tau \in \{1, 17/16, 9/8, 5/4, 3/2, 5\}$ . Setting  $\tau = 1$  corresponds to the approximation implied by the Whittle likelihood (Whittle, 1954). Choosing  $\tau = 5$  is intended to show how  $\lambda^\tau$  behaves when  $R$  is a very good approximation to  $K$ . In Table 1, we present the results of the first numerical study. When  $\tau = 1$ ,  $\lambda^\tau < \lambda_0$ , although  $\lambda^\tau$  increases with  $n$ . This is not surprising because we expect the Whittle likelihood to underestimate range parameters since the Whittle likelihood assumes periodic correlation on  $\delta\mathbb{J}_n$  when the true covariance function is not periodic at all. For every  $n$ ,  $\lambda^\tau$  approaches  $\lambda_0$  as  $\tau$  increases, with  $\lambda^\tau$  converging more quickly for larger  $n$ . For every  $\tau > 1$ ,  $\lambda^\tau$  approaches  $\lambda_0$  as  $n$  increases. This last remark is an important one because it suggests that we obtain very accurate approximations with small  $\tau$  when the number of ob-

$\tau$	$n$			
	$32^2$	$48^2$	$64^2$	$80^2$
1	0.1234	0.1310	0.1353	0.1380
17/16	0.1457	0.1484	0.1493	0.1496
9/8	0.1485	0.1495	0.1498	0.1499
5/4	0.1496	0.1499	0.1500	0.1500
3/2	0.1499	0.1500	0.1500	0.1500
5	0.1500	0.1500	0.1500	0.1500

Table 1: Numerical values of  $\lambda^\tau$  for various choices of  $\tau$  and  $n$  with constant lattice spacing  $(32\sqrt{2})^{-1}$  and  $\lambda_0 = 0.15$ .

$\tau$	$\lambda_0$				
	0.05	0.10	0.15	0.20	0.25
1	0.0482	0.0932	0.1353	0.1748	0.2120
17/16	0.0500	0.0997	0.1493	0.1986	0.2477
9/8	0.0500	0.0999	0.1498	0.1996	0.2494
5/4	0.0500	0.1000	0.1500	0.1999	0.2499
3/2	0.0500	0.1000	0.1500	0.2000	0.2500
5	0.0500	0.1000	0.1500	0.2000	0.2500

Table 2: Numerical values of  $\lambda^\tau$  for various choices of  $\tau$  and  $\lambda_0$  for a lattice with  $64^2$  points and spacing  $(32\sqrt{2})^{-1}$ . The third column of this table is equivalent to the third column of Table 1

servations is large, which is desirable because the methods are designed for analyzing very large datasets.

The second numerical study considers the behavior of  $\lambda^\tau$  for various choices of  $\lambda_0$  to understand how the performance of the approximation depends on the strength of the spatial correlation. We fix the lattice spacing at  $\delta = (32\sqrt{2})^{-1}$  and the number of lattice locations at  $64^2$ . We vary  $\lambda_0 \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$ . The results presented in Table 2 show that  $\lambda^\tau$  approaches  $\lambda_0$  as  $\tau$  increases for every choice of  $\lambda_0$ . We also observe that  $\lambda^\tau$  converges faster to  $\lambda_0$  when  $\lambda_0$  is small, that is, when the spatial correlation is weak. In every case,  $\tau = 3/2$  is large enough to ensure that  $\lambda^\tau$  and  $\lambda_0$  agree to four decimal places. In Figure 2, we plot the target covariance functions and several of the approximations with  $\lambda = 0.25$ . It is important to note that approximations need not be accurate at all spatial lags in order for  $\lambda^\tau$  to be very close to  $\lambda_0$ . To say this more concretely with a specific example, when  $\lambda_0 = 0.25$  and  $\tau = 5/4$ ,  $\lambda^\tau = 0.2499$  even though  $R_{\mathbf{m}}(\delta(63, 0)) = 0.2283$  is not close to the target covariance  $K_{\lambda_0}(\delta(63, 0)) = 0.0038$ . This suggests that, in this specific example, it is not necessary for  $R_{\mathbf{m}}(\mathbf{h})$  to well approximate  $K(\mathbf{h})$  at large lags in order to produce a likelihood function that returns accurate parameter estimates.

The third numerical study addresses how the approximations perform in the fixed

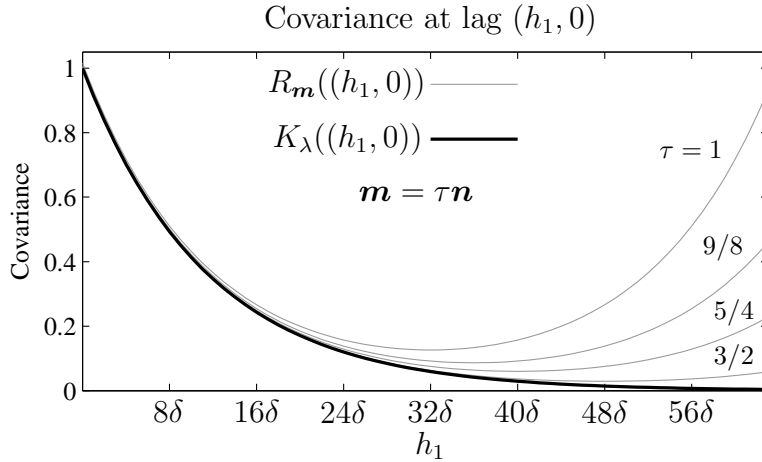


Figure 2: Target covariance function and approximations with several values of  $\tau$  for  $\lambda = 0.25$ .

$\tau$	$n$			
	$32^2$	$48^2$	$64^2$	$80^2$
1	0.1234	0.1235	0.1236	0.1237
17/16	0.1457	0.1474	0.1483	0.1488
9/8	0.1485	0.1492	0.1495	0.1497
5/4	0.1496	0.1498	0.1499	0.1499
3/2	0.1499	0.1500	0.1500	0.1500
5	0.1500	0.1500	0.1500	0.1500

Table 3: Numerical values of  $\lambda^\tau$  for various choices of  $\tau$  and  $n$  when the resolution of the lattice increases on a fixed domain. The lattice spacing is  $(\sqrt{n}\sqrt{2})^{-1}$ , and  $\lambda_0 = 0.15$ . The first column of this table is the same as the first column of Table 1.

domain scenario, when the size of the spatial domain is held constant, and the lattice spacing decreases. Stein (1999) shows that in this setting, one can consistently estimate the local properties of the covariance function. Here,  $\lambda$  controls the derivative of the covariance function at the origin. We use lattice spacing  $(\sqrt{n}\sqrt{2})^{-1}$  with  $n \in \{32^2, 48^2, 64^2, 80^2\}$ , and we set  $\lambda_0 = 0.15$ . The results are reported in Table 3. When  $\tau = 1$ ,  $\lambda^\tau$  is again less than  $\lambda_0$ , but in this fixed domain scenario,  $\lambda^1$  does not improve much as  $n$  increases; it changes from  $\lambda^1 = .1234$  when  $n = 32^2$  to  $\lambda^1 = .1237$  when  $n = 80^2$ . When  $\tau > 1$ , however,  $\lambda^\tau$  does appear to be approaching  $\lambda_0$  as  $n$  increases, even when  $\tau$  is as small as 17/16. In that case  $\lambda^{17/16} = 0.1457$  when  $n = 32^2$  versus  $\lambda^{17/16} = 0.1488$  when  $n = 80^2$ . The error  $\lambda^{17/16} - \lambda_0$  decreases by 72% as we increase  $n$  from  $32^2$  to  $80^2$  versus a 1.3% decrease in the error  $\lambda^1 - \lambda_0$  for the same increase in  $n$ . As before,  $\lambda^\tau$  approaches  $\lambda_0$  as  $\tau$  increases in every case.



## 4 Simulations

This section presents simulations that provide further support that we can obtain very accurate parameter estimates with small values of  $\tau$ . Choosing a small value of  $\tau$  defines a relatively small embedding lattice, which we show has a dramatic effect on the speed of convergence of the Monte Carlo inferential procedures. With small embedding lattices, the number of imputed values is not large relative to the number of observations, so the observations hold greater authority in driving the parameter updates in the iterative algorithms. Using a large embedding lattice leads to highly correlated Markov chains and slowly converging Monte Carlo EM algorithms. Standard circulant embedding requires an embedding lattice that is at least two times larger—and in practice often three or more times larger—than the observation lattice in each dimension. We show that using approximate covariances allows us to obtain accurate parameter estimates with  $\tau$  as small as 1.25.

To demonstrate these points, we focus our simulation studies on the estimation of a single parameter in the powered exponential covariance function. In Section 5, we pursue the estimation of multiple parameters in the powered exponential covariance function, as well as in the Matérn covariance function for the photosynthetically active radiation data. The isotropic powered exponential covariance is a flexible and commonly used covariance function defined by

$$K(\mathbf{h}) = \sigma^2 \exp\left(-(\|\mathbf{h}\|/\lambda)^\alpha\right) + \gamma \mathbf{1}(\mathbf{h} = 0),$$

where  $\sigma^2, \lambda, \gamma > 0$  and  $\alpha \in (0, 2]$ . We simulate 100 spatial data sets on a lattice of size  $\mathbf{n} = (64, 64)$  with spacing  $(64\sqrt{2})^{-1}$  from a mean-zero Gaussian process model with powered exponential covariance function with  $\sigma^2 = 4$ ,  $\lambda = 0.1$ ,  $\alpha = 1$  (exponential model), and  $\gamma = 0.01$ . We focus on the estimation of  $\alpha$  and specify its prior to be uniform over  $(0, 2]$ . The lattice dimensions, covariance function and parameters, and prior are the same as those used in a simulation Stroud et al. (2014). Here, all parameters except for  $\alpha$  are assumed to be known and fixed at their true values.

The MCMC procedure consists of Metropolis-Hastings updates of  $\alpha$ , where the acceptance probabilities are tuned to 0.50 during a burn-in period of 1000 iterations. Circulant embedding with the exact model is achieved with the same procedures outlined in Stroud et al. (2014), which give an embedding lattice of size  $\mathbf{m} = (192, 192)$ . Circulant embedding with approximate covariances is carried out using the methods in Section 2 with choices of expansion parameter  $\tau = 5/4$  and  $3/2$ , giving embedding lattices of size  $\mathbf{m} = (80, 80)$  and  $(96, 96)$ . We compute  $R_{\mathbf{m}}(\mathbf{h})$  using a truncation of the wrapping of  $K(\cdot)$  with  $N = 3$ , which always produced positive definite covariance matrices.

The lattice contains  $n = 4092$  observation locations, so for comparison we can also numerically evaluate the posterior in order to obtain posterior mean  $\hat{\alpha}_j$  and posterior standard error  $\text{se}(\hat{\alpha}_j)$  for each simulation  $j$ . For each of the MCMC runs, after a burn-in period of 1000 iterations, we calculate  $\tilde{\alpha}_{j,k}$ , a running mean after  $k$  samples

of  $\alpha$  for the  $j$ th simulation. In order to assess the accuracy and the convergence of the MCMC estimate, we run each chain for 40,000 post-burn-in iterations, and record

$$k_j(\varepsilon) = \max \left\{ k : \frac{|\tilde{\alpha}_{j,k} - \hat{\alpha}_j|}{\text{se}(\hat{\alpha}_j)} > \varepsilon \right\},$$

the last iteration for which the standardized absolute difference between the running MCMC mean and the posterior mean is larger than  $\varepsilon$ . We consider  $\varepsilon \in \{0.05, 0.1, 0.2\}$ . We also record the time required to reach  $k_j(\varepsilon)$ . This allows us to study the number of iterations and time required to reach a desired level of accuracy. All computations are completed with a machine running Matlab R2013a on an Intel Core i7-4770 processor at 3.40GHz with 32GB of RAM. In Figure 3, we plot the empirical cumulative distribution function of  $k_j(\varepsilon)$  for each choice of  $\varepsilon$ , and for each of the three estimation methods, i.e. embedding with approximate covariances with  $\tau = 5/4$  and  $3/2$  and embedding with exact covariances.

We see that the approximate methods outperform the exact method at every tolerance  $\varepsilon$ . Over 96% of the Markov chains using the approximation  $\tau = 5/4$  had converged to tolerance  $\varepsilon = 0.05$  within 20000 iterations, whereas only 51% of the exact chains had converged to within the same tolerance by 20000 iterations. Put another way, 89% of the chains using the exact covariances had converged to within tolerance 0.05 after 40000 iterations, whereas it required just 13,500 iterations for 89% of the chains to converge when the  $\tau = 5/4$  approximation was used. We see similar advantages for the larger tolerances as well. In Figure 4, we show the time required for the chains to reach desired levels of tolerance. For example, the first plot shows that after running the chains for 30 minutes, 97% of the chains using the approximation  $\tau = 5/4$  had converged to tolerance 0.05, while just 50% of the chains using exact covariances had converged. For the weaker tolerance  $\varepsilon = 0.1$ , 93% of the exact covariance chains had converged after 30 minutes, whereas 100% of the approximate covariance chains had converged after just 9 minutes.

Using the same 100 simulated datasets, we implement the Monte Carlo EM algorithm proposed in Stroud et al. (2014). The Monte Carlo EM algorithm specifies  $M$ , the number of conditional simulations over which the loglikelihood is averaged in each iteration of the algorithm. To see the effect of the choice of  $M$ , we use  $M = 20$  and  $M = 100$  to analyze each simulated data set. Since this is a Monte Carlo EM algorithm, the parameter iterates do not converge to any particular value. For this reason, we suggest estimating parameters by averaging the parameter iterates after a “burn-in” period has concluded. We also numerically evaluate the exact likelihood to obtain maximum likelihood parameter estimates. Thus we can compare root mean squared differences between the maximum likelihood estimates and the Monte Carlo EM estimates found using various choices of burn-in iterations and averaging iterations. The results for cutoff embedding and our approximate procedures with  $\tau = 1.25$  are plotted in Figure 5. We see that the Monte Carlo EM algorithm with cutoff embedding requires more burn-in iterations for the estimates to stabilize. Further, even

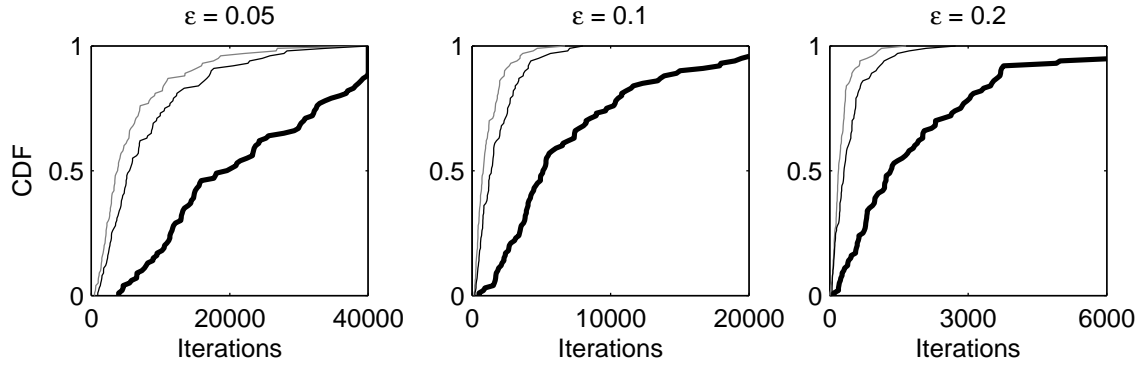


Figure 3: Empirical CDF of  $k_j(\varepsilon)$  for three choices of  $\varepsilon$ , and three estimation methods: approximate with  $\tau = 5/4$  (thin gray), approximate with  $\tau = 3/2$  (thin black), exact (thick black).

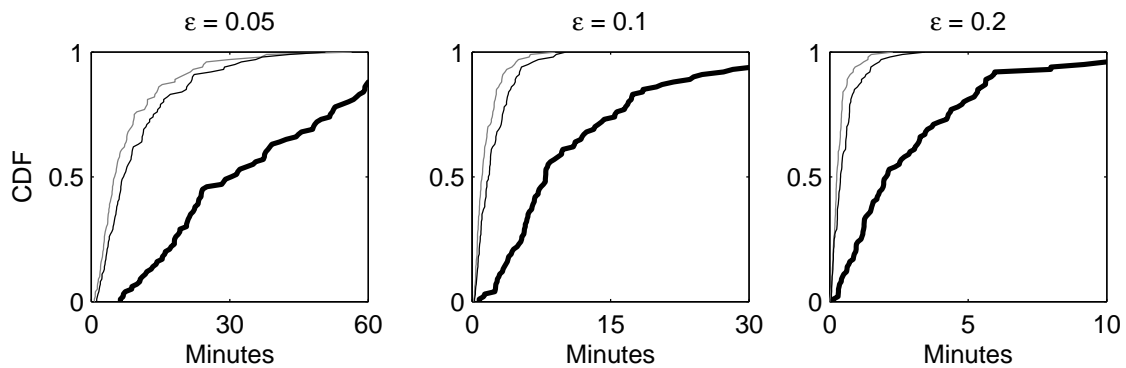


Figure 4: Empirical CDF for time required to reach  $k_j(\varepsilon)$  for three choices of  $\varepsilon$ , and three estimation methods: approximate with  $\tau = 5/4$  (thin gray), approximate with  $\tau = 3/2$  (thin black), exact (thick black).

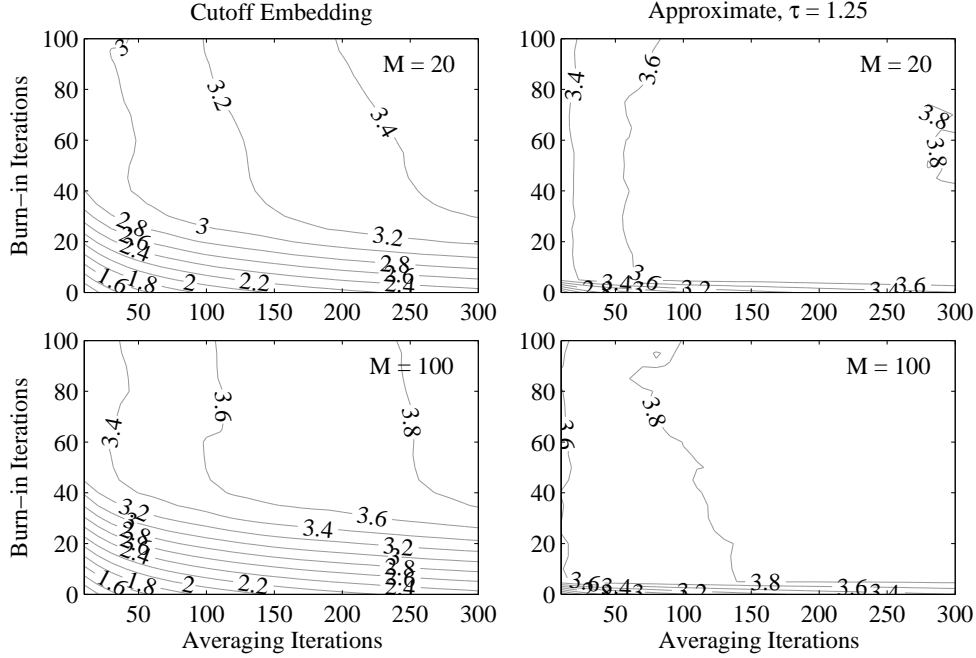


Figure 5: Contour plots of  $-\log_{10}$  of root mean squared deviations between Monte Carlo EM estimates and maximum likelihood estimates of  $\alpha$ , where the mean squared differences are taken over the 100 simulated datasets. The approximate covariances use  $\tau = 1.25$ .

if we set the burn-in time to 100 iterations, the estimates found using approximate covariances converge faster to the maximum likelihood estimates; when  $M = 20$ , the approximations need fewer than 30 averaging iterations in order for the root mean squared differences to fall below  $10^{-3.4} \approx 0.0004$ , whereas cutoff embedding needs 200 iterations to reach this tolerance. Thus, our approximate procedures provide remarkable reductions in the number of iterations required for both burn-in and averaging in the Monte Carlo EM algorithm.

## 5 Photosynthetically Active Radiation Data

Aqua is a NASA satellite mission whose central aim is to collect information about Earth’s water cycle. As is typical of most polar-orbiting satellites, Aqua’s measurements do not attain complete global coverage on short time scales; a typical daily map of Aqua data contains large swaths of missing values at locations over which Aqua did not orbit. Our goal in this section is to provide complete spatial maps of Photosynthetically Active Radiation (PAR) over a region for which there are a substantial number of missing values. PAR, which is detected by the Moderate Resolution Imaging Spectrometer (MODIS), quantifies the abundance of light at wavelengths between

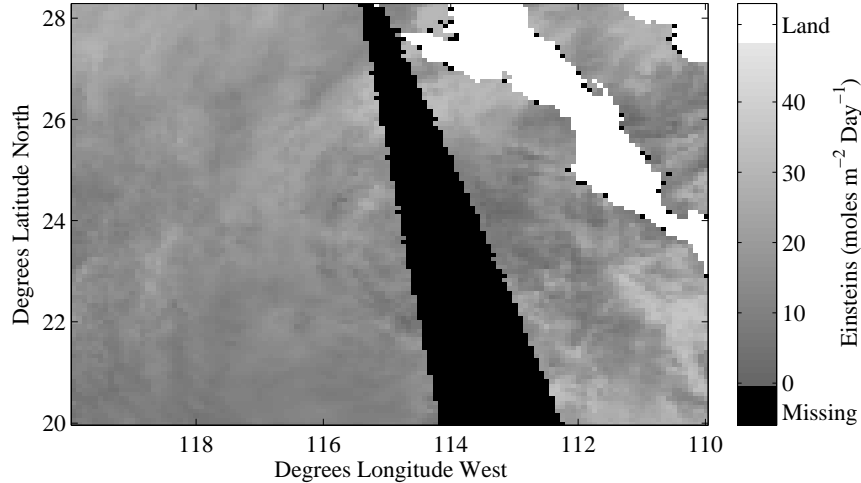


Figure 6: Grayscale map of photosynthetically active radiation from December 1, 2013. Missing values are black, and land pixels are white; the diagonal strip of land is Mexico’s Baja California peninsula. The resolution is  $1/12^\circ$  in both latitude and longitude.

400 and 700 nm, the spectral range of radiation that organisms use in photosynthesis, and thus is an important quantity affecting biological systems. In Figure 6, we plot a map of a daily gridded data product of PAR values located west of Mexico’s Baja California peninsula. The data can be downloaded from <http://oceancolor.gsfc.nasa.gov>, and this particular dataset is from December 1, 2013. PAR values derived from Aqua’s measurements are reported only over the oceans. There is a triangular region of missing observations, as well as a few missing along the coasts. We aim to interpolate the missing observations with values that match the statistical properties of the observed process to obtain physically plausible reconstructions of PAR. To accomplish this, we use the conditional simulations of the missing values that are required as part of the computationally efficient estimation methods presented in this paper, and we report an ensemble of the conditional simulations to provide accurate indications of the uncertainty in the interpolations.

The PAR lattice presented in Figure 6 contains 120 evenly-spaced longitude values and 100 evenly-spaced latitude values at a resolution of  $1/12^\circ$  in both latitude and longitude, for a total of 12,000 lattice locations. There are 2,412 lattice locations for which PAR is missing, due either to the pixel being a land pixel or the value being genuinely missing, giving 9,588 observed PAR values. The data do not possess any obvious deviations from the isotropic Gaussian assumption, nor are there any discernible trends in the data. After subtracting the empirical mean of the observations, we consider three covariance models for PAR anomalies: (1) mean-zero isotropic powered exponential covariance with zero nugget and unknown  $(\sigma^2, \lambda, \alpha)$ , (2) mean-zero

isotropic Matérn covariance with zero nugget,

$$K_{\theta}(\mathbf{h}) = \sigma^2 \frac{(\|\mathbf{h}\|/\lambda)^{\nu} \mathcal{K}_{\nu}(\|\mathbf{h}\|/\lambda)}{2^{\nu-1} \Gamma(\nu)},$$

and unknown  $(\sigma^2, \lambda, \nu)$ , and (3) a mean-zero model with aliased spectral density

$$f_{\delta}(\boldsymbol{\omega}) = \sigma^2 c_{\lambda, \nu} \left[ 1 + \left( \frac{\lambda}{\delta} \right)^2 \left( \sin^2 \left( \frac{\delta \omega_1}{2} \right) + \sin^2 \left( \frac{\delta \omega_2}{2} \right) \right) \right]^{-\nu-1}. \quad (5)$$

We refer to the model in (5) as the quasi Matérn model due to its similarity to the Matérn spectral density and give an asymptotic justification for this name in Appendix B of the Supplementary Material, along with a more general specification. The parameters  $(\sigma^2, \lambda, \nu)$  have the same interpretations as in the Matérn model. The coefficient  $c_{\lambda, \nu}$  is a normalizing constant, computed numerically. The quasi Matérn is defined in terms of its aliased spectral density, and thus no wrapping of covariances or spectral densities is required for the computations described in this paper; we simply evaluate  $f_{\delta}(\boldsymbol{\omega})$  at the Fourier frequencies associated with  $\mathbf{m}$  and transform the resulting array of spectral density values with an inverse FFT to obtain the associated covariances. We assume  $(\sigma^2, \lambda, \nu)$  are unknown.

We implement the MCMC methods to estimate the parameters in all models. To simplify notation across the models, we define  $\boldsymbol{\theta} = (\lambda, \alpha)$  if the model is the powered exponential or  $\boldsymbol{\theta} = (\lambda, \nu)$  if the model is either the Matérn or quasi Matérn. We specify prior  $\pi(\sigma^2, \boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta})/\sigma^2$ , where  $\pi(\boldsymbol{\theta}) = 1/2(1 + \lambda/2)^{-2}$  in the powered exponential model, which places a uniform prior on  $\alpha$  over  $(0, 2)$ , and  $\pi(\boldsymbol{\theta}) = 1/4(1 + \lambda/2)^{-2}(1 + \nu/2)^{-2}$  in the Matérn and quasi Matérn models. We update  $\boldsymbol{\theta}$  with a MH algorithm with a bivariate normal proposal distribution on the log scale. The full conditional  $\pi(\sigma^2 | \boldsymbol{\theta}, \mathbf{Z})$  is inverse gamma  $IG((m-1)/2, S^2(\boldsymbol{\theta})/2)$ , where  $S^2(\boldsymbol{\theta}) = \mathbf{Z}'C(\boldsymbol{\theta})^{-1}\mathbf{Z}$ , and  $C(\boldsymbol{\theta})$  is the correlation matrix corresponding to parameter vector  $\boldsymbol{\theta}$ . This is the standard conjugate family for variance parameter  $\sigma^2$ . The bivariate lognormal proposal distribution for  $\lambda$  and  $\nu$  is tuned to have acceptance probability of 0.5 during 5,000 burn-in iterations. In cutoff embedding, we use an embedding lattice of size  $\mathbf{m} = (336, 336)$ , which is highly composite and has 112,876 lattice locations. Our approximate methods are implemented with  $\tau = 1.25$ , which gives embedding lattices of size  $\mathbf{m} = (125, 150)$  and  $m = 18,750$  total lattice locations. We wrap the covariances with  $N = 2$ , which always resulted in positive definite covariance matrices, and no wrapping is required for the quasi Matérn covariance.

The multiple spatial covariance parameters, especially the variance and range parameters, tend to be highly correlated in this example, and thus the MCMC algorithms return highly correlated parameter iterates, even though we can sample from the full conditional distribution of the variance parameter. Since the most time-consuming step in the algorithm is conditionally simulating the missing data on the embedding lattice, we propose taking several MH steps of the covariance parameters

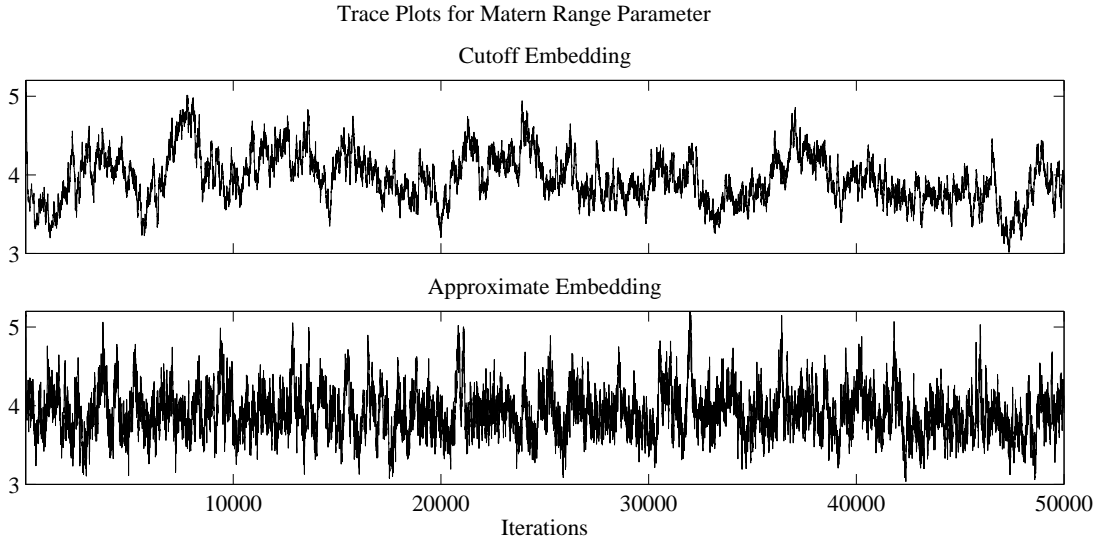


Figure 7: Trace plots for Matérn range parameter from the PAR data, using cutoff embedding and approximate embedding with  $\tau = 1.25$

per conditional simulation, and we obtained good results when we took four MH steps per conditional simulation. The Markov chains are each run for 50,000 iterations. We provide trace plots for the Matérn range parameter in Figure 7, showing an example of the smaller autocorrelation in the chain that uses approximate embedding, which is typical of all parameters and consistent with the results in the simulation study. The lag 10 autocorrelations for the Matérn range parameters are 0.967 for cutoff embedding and 0.807 for approximate embedding. With the help of Matlab’s profiling capabilities, we record the CPU time per iteration attributed to the various computational tasks required for the MCMC algorithms. We report those results in Figure 8. The three most time-consuming tasks are the conditional simulations, the loglikelihoods required for evaluating the MH acceptance probabilities, and the construction of the covariance arrays. The approximate methods are faster overall than cutoff embedding, which required 0.88 seconds per iteration with the powered exponential covariance and 0.87 seconds per iteration with the Matérn covariance, whereas the approximate methods required 0.29 seconds and 0.40 seconds per iteration for the powered exponential and the Matérn models. The approximate methods devoted a significant amount of time—0.19 seconds per iteration for the Matérn—to constructing the periodic covariance arrays, a consequence of the wrapping of the covariances. The approximate methods for the quasi Matérn devote a negligible amount of time to constructing the covariance arrays and is the fastest overall of the three methods, taking 0.22 seconds per iteration, nearly four times faster than the Matérn with cutoff embedding. All methods use a preconditioner corresponding to the submatrix of the complete data precision matrix and require roughly 50 iterations for the precondi-

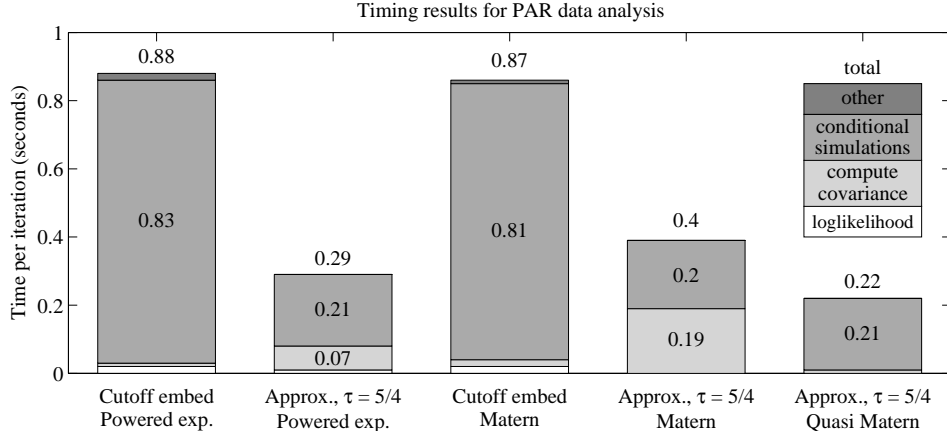


Figure 8: CPU time required per MCMC iteration with various models and embedding methods, separated by three main computational tasks.

tioned conjugate gradient algorithms to converge on average. We experimented with several forms of a preconditioner based on the Stein et al. (2004) likelihood approximation, which can be made to converge in a smaller number of iterations but was slower overall in this instance.

We now investigate the quality of the fitted models in terms of exact loglikelihood. There are several ways to construct fitted models from the Markov chains, one of which is to compute  $K_{\hat{\theta}}(\mathbf{h})$ , where  $\hat{\theta}$  as an average of the parameter iterates in the chain. We refer to this as the *averaged parameter* fitted model. For this estimate, we thin the chain, taking only every tenth iterate of the 45,000 post-burn-in iterations for the average. We also construct separate fitted models by averaging the covariances associated with the parameter iterates. Specifically, we compute

$$\hat{K}(\mathbf{h}) = \frac{1}{B} \sum_{i=1}^B K_{\theta^i}(\mathbf{h}), \quad (6)$$

where  $\theta^i$  is the  $i$ th thinned MCMC sample, with  $B = 4500$  samples. We refer to the estimate in (6) as the *averaged covariance* fitted model. In cases where the loglikelihood has irregularly shaped contours, the averaged parameter model could differ substantially from the averaged covariances model. We note that in constructing the covariance estimates, we always use the covariance function  $K_{\theta}$ , not  $R_{\mathbf{m}}$ , because the approximations are not guaranteed to accurately reflect the target covariance at large spatial lags (see Figure 2). When using the quasi Matérn model we approximate  $K_{\theta}(\mathbf{h})$  by discretizing the integral in (2) over a very fine grid with  $\mathbf{m} = 4\mathbf{n}$ , which is still efficient to compute with FFT algorithms. In Table 4, we include the exact loglikelihood values of the various model estimates. The Matérn and quasi Matérn models provide better fits than the powered exponential model in terms of loglikelihood. All of the fitted Matérn and quasi Matérn models agree to within a few loglikelihood units,



Model	Method	Loglikelihood	
		Averaged Covariance	Averaged Parameter
Powered exp.	Cutoff Embedding	-52.66	-53.16
Powered exp.	Approximate, $\tau = 5/4$	-51.32	-52.04
Matérn	Cutoff Embedding	0	-0.57
Matérn	Approximate, $\tau = 5/4$	-0.03	-0.54
quasi Matérn	Approximate, $\tau = 5/4$	-1.14	-1.65

Table 4: Table of loglikelihoods for the three models and two methods. We compute the exact Gaussian loglikelihoods for the model estimate constructed by averaging the covariances and by averaging the parameters. Loglikelihood differences from that of the Matérn model fit with the approximate method are reported.

which is negligible for a dataset of this size. It is actually quite remarkable that even though the quasi Matérn model is not equivalent to the Matérn model, in the sense that the aliased spectral density of the Matérn is not equal to the spectral density in (5), both fitted models give nearly the same loglikelihoods, an indication of the flexibility of the two models. Averaging covariances provided slightly better fits than averaging parameters for every model. The models obtained by approximate methods are roughly equal in terms of loglikelihood to the corresponding model estimate obtained by using cutoff embedding, and they give roughly equal posterior mean parameter estimates—for example the Matérn smoothness parameters are estimated to be 0.8578 (s.d. 0.02) with cutoff embedding and 0.8595 (s.d. 0.02) with approximate embedding. Thus, the approximations provided computational benefits without any sacrifice in the quality of the fitted models they produced.

Finally, in Figure 9, we plot conditional simulations of the PAR process over the ocean pixels of the observation region. The three conditional simulations use three different sets of parameters taken from the quasi Matérn Markov chain at iterations 10,000, 20,000, and 30,000, so the conditional simulations incorporate the uncertainty of the parameters. The conditional simulations produce PAR values over the land pixels as well, but we do not plot those since they are not reported in Aqua MODIS datasets. The PAR values in the three conditional simulations are exactly the same except for a few pixels along the coasts and the pixels in the triangular swath indicated by the thin black lines, which merge neatly with the observed values on the borders of the swath. The interpolated values also match the second-order properties of the rest of the dataset because they are simulated from a covariance model that is fit to the observed data. In scientific applications where it is necessary to have a complete map of PAR values as an input into larger model, the three (or possibly more) conditional simulations could be used to propagate the uncertainty associated with the interpolations and the fitted spatial model through the analysis. The methods discussed in this paper offer a way to produce an ensemble of complete interpolated maps in a computationally efficient manner.

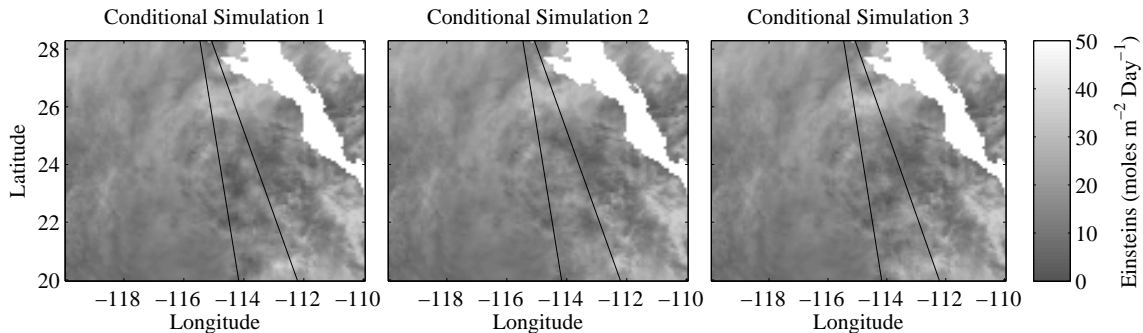


Figure 9: Three conditional simulations with Quasi Matérn covariance parameters taken from MCMC chain at iterations 10,000, 20,000, and 30,000.

## 6 Discussion

Numerical methods based on circulant embedding of covariance matrices are powerful tools for statistical computations involving lattice data and stationary models. Not only do they avoid the  $O(n^3)$  flops required for Cholesky decompositions of covariance matrices, they circumvent the need to store the  $O(n^2)$  covariance matrices. The use of circulant embedding for simulation of stationary Gaussian processes on regular lattices has a mature history, and the recent work of Stroud et al. (2014) is an important step forward in uncovering ways to exploit circulant embedding for making inference from lattice data.

We demonstrate that the discrete spectral approximation can be advantageously employed in order to reduce the size of the embedding lattice, while preserving periodicity on the embedding lattice, positive definiteness, and the ability to exploit fast computational algorithms. This approximation allows the user to control its accuracy via an expansion parameter, and our numerical studies show that, with expansion parameters as small as 1.25 or 1.5, the bias introduced by this approximation is negligible among a range of commonly encountered scenarios of spatial correlation. The size of the bias also decreases with the size of the dataset, both in increasing domain and fixed domain scenarios. While the approximations are accurate at short spatial lags and give rise to likelihood functions that perform well for estimating parameters, the approximation to individual covariances may not be accurate at large spatial lags (see Figure 2), and thus we recommend using the target covariance function with the estimated parameters to characterize covariances. We also provide novel methods, along with the necessary theoretical support, for computing the discrete spectral approximation when the covariance function is specified.

Our simulation studies demonstrate the importance of reducing the size of the embedding lattice. When cutoff embedding is employed, the embedding lattice is at least four times the size of the observation lattice, and the particular version described in Stroud et al. (2014) requires an embedding lattice at least nine times

the size of the observation lattice. We found that both MCMC and Monte Carlo EM parameter estimates are slow to converge when cutoff embedding is employed. The use of approximate methods that decrease the size of the embedding lattice substantially improves the convergence of the parameter estimates, over a range of accuracy tolerances, as compared to numerically evaluated posterior means.

Extremely long correlation ranges, very smooth data, or small observation lattices may require expansion parameters larger than 1.5. However, we expect that in situations where the spatial correlation is very strong, cutoff embedding with  $m = 3n$  will likely not produce positive definite covariance arrays. We expect that the embedding lattice required for cutoff embedding to produce positive definite covariance arrays will generally always be larger than the lattice required for the approximate methods to produce extremely sharp approximations. Making the previous conjecture more precise is an avenue for future work. When analyzing very strongly correlated spatial data, we suggest conducting a sensitivity analysis with a few values of  $\tau \in [1.25, 2]$ , with the reminder that larger values of  $\tau$  will require more iterations to achieve convergence.

We implemented the MCMC algorithms to analyze satellite observations of photosynthetically active radiation. Three spatial models were considered, each of them described by three parameters, and we used both cutoff embedding and the discrete spectral approximation in estimating them. Based on our analysis, we suggest using multiple MH steps per conditional simulation in order to improve the MCMC algorithm, since the parameter iterates tended to be highly correlated, and the likelihood given the full dataset on the embedding lattice is cheap to evaluate compared to the cost of the conditional simulations. Although we did not conduct a formal analysis of the trace plots, the use of the discrete spectral approximation improved their appearance compared to those obtained using cutoff embedding, which is consistent with the results of the simulation study. Further, the computation time per iteration of the MCMC algorithms was smaller with the use of the approximate methods, and the speedup was greatest when the aliased spectral density was modeled directly, since this approach does not require spectral wrapping or evaluation of the Matérn covariance function. Based on these results, we strongly recommend modeling the covariances in the spectral domain with an aliased spectral density, and we suggest the quasi Matérn as a flexible model.

Several aspects of the material presented here can be easily generalized. The powered exponential and Matérn covariance functions are isotropic, but we do not require isotropy or even geometric anisotropy, only stationarity. The quasi Matérn model has a generalization to  $d$  dimensions, which we present in Appendix B of the Supplementary Material. We also assume that the lattice has equal spacing in every dimension, but the models and methods are easily generalized to situations with different spacing  $\delta_j$  in each dimension  $j$ . In this case, the aliased spectral densities are defined on  $\prod_{j=1}^d [0, 2\pi/\delta_j]$ . The analyses that we presented used a common expansion factor for each dimension, but this is not required, and one can see how it may be com-

putationally advantageous to use a smaller expansion factor in a dimension that is very large or has weak correlation along that dimension. The methods are applicable for multivariate spatial data as well. Guinness et al. (2014) provide a framework for defining multivariate spatial lattice models in the spectral domain.

Appendices and Matlab code for reproducing the results appearing in this paper are available as online supplementary material.

## Acknowledgements

The authors are supported by the National Science Foundation’s Research Network for Statistical Methods for Atmospheric and Oceanic Sciences, award number 1107046, and National Science Foundation Award number 1406016.

## A Nested Block Circulant Matrices

A matrix is block circulant if it can be written in block form as

$$K = \begin{bmatrix} K_0 & K_1 & K_2 & \cdots & K_{n-1} \\ K_{n-1} & K_0 & K_1 & \cdots & K_{n-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ K_1 & K_2 & K_3 & \cdots & K_0 \end{bmatrix},$$

where the blocks are all the same size and may be of size 1, in which case we also say that the matrix is circulant. Nested block circulant matrices are defined recursively: a matrix is nested block circulant if it is block circulant, and each subblock  $K_j$  is also nested block circulant. Covariance matrices have the additional properties that they are symmetric and positive definite. Symmetry implies that  $K_0$  is symmetric, and  $K_{n-j} = K_j'$ .

## B Quasi Matérn covariance

The quasi Matérn model presented in Section 5 has a generalization to  $d$  dimensions. In this case, the spectral densities are defined on  $[-\pi/\delta, \pi/\delta]^d$ , and the model is

$$f_\delta(\boldsymbol{\omega}) = \sigma^2 \left( 1 + \left( \frac{\alpha}{\delta} \right)^2 \left( \sum_{j=1}^d \sin^2 \left( \frac{\delta \omega_j}{2} \right) \right) \right)^{-\nu-d/2}.$$

This expression converges pointwise as  $\delta \rightarrow 0$  to

$$\sigma^2 \left( 1 + \left( \frac{\alpha}{2} \right)^2 \|\boldsymbol{\omega}\|^2 \right)^{-\nu-d/2},$$

which is one parametric form for the spectral density of the isotropic Matérn covariance function, justifying the name quasi Matérn.

## C Proof

The following lemma is a multidimensional version of the Poisson summation formula and is of particular use for approximating the covariance functions introduced in Section 2.

**Lemma C.1.** *If  $f_\delta$  is the continuous aliased spectral density for  $K$ , then  $R_{\mathbf{m}}(\mathbf{h}) = \sum_{\mathbf{j} \in \mathbb{Z}^d} K(\mathbf{h} + \delta \mathbf{j} \circ \mathbf{m})$ , where  $\mathbf{j} \circ \mathbf{m} = (j_1 m_1, \dots, j_d m_d)$ .*

To simplify the notation, we set  $\delta = 1$  in the proof. It is no more difficult to prove with arbitrary  $\delta$ , only more cumbersome notationally.

*Proof.* We write:

$$\begin{aligned} & \sum_{j_1=-N}^N \cdots \sum_{j_d=-N}^N K((h_1 + j_1 m_1, \dots, h_d + j_d m_d)) \\ &= \sum_{j_1=-N}^N \cdots \sum_{j_d=-N}^N \int_{[0, 2\pi]^d} f_1(\boldsymbol{\omega}) e^{i\boldsymbol{\omega}'\mathbf{h}} e^{i\omega_1 j_1 m_1} \cdots e^{i\omega_d j_d m_d} d\boldsymbol{\omega} \\ &= \int_{[0, 2\pi]^d} f_1(\boldsymbol{\omega}) e^{i\boldsymbol{\omega}'\mathbf{h}} \left( \sum_{j_1=-N}^N e^{i\omega_1 j_1 m_1} \right) \cdots \left( \sum_{j_d=-N}^N e^{i\omega_d j_d m_d} \right) d\boldsymbol{\omega}. \quad (7) \end{aligned}$$

Since the integrand in (7) is periodic in each dimension, integrating over  $[0, 2\pi]^d$  is equivalent to integrating over

$$\prod_{k=1}^d \left[ -\frac{\pi}{m_k}, 2\pi - \frac{\pi}{m_k} \right] = \bigcup_{\boldsymbol{\ell} \in \mathbb{J}_{\mathbf{m}}} \prod_{k=1}^d \left[ \frac{2\pi}{m_k}(\ell_k - 1/2), \frac{2\pi}{m_k}(\ell_k + 1/2) \right] \stackrel{\text{def.}}{=} \bigcup_{\boldsymbol{\ell} \in \mathbb{J}_{\mathbf{m}}} A_{\boldsymbol{\ell}},$$

so that the integral in (7) can be written as

$$\sum_{\boldsymbol{\ell} \in \mathbb{J}_{\mathbf{m}}} \int_{A_{\boldsymbol{\ell}}} f_1(\boldsymbol{\omega}) e^{i\boldsymbol{\omega}'\mathbf{h}} \left( \sum_{j_1=-N}^N e^{i\omega_1 j_1 m_1} \right) \cdots \left( \sum_{j_d=-N}^N e^{i\omega_d j_d m_d} \right) d\boldsymbol{\omega}.$$

The quantities in parentheses converge to periodic delta functions with period  $2\pi/m_k$  in  $\omega_k$  (? , Section 1.17(iii)). Since  $f_1$  is continuous, the integral of the integrand  $f_1(\boldsymbol{\omega}) e^{i\boldsymbol{\omega}'\mathbf{h}}$  with respect to the periodic delta functions is equal to the integrand evaluated at the Fourier frequencies,

$$\begin{aligned} & \lim_{N \rightarrow \infty} \sum_{\boldsymbol{\ell} \in \mathbb{J}_{\mathbf{m}}} \int_{A_{\boldsymbol{\ell}}} f_1(\boldsymbol{\omega}) e^{i\boldsymbol{\omega}'\mathbf{h}} \left( \sum_{j_1=-N}^N e^{i\omega_1 j_1 m_1} \right) \cdots \left( \sum_{j_d=-N}^N e^{i\omega_d j_d m_d} \right) d\boldsymbol{\omega} \\ &= \frac{(2\pi)^d}{m} \sum_{\boldsymbol{\ell} \in \mathbb{J}_{\mathbf{m}}} f_1(\boldsymbol{\omega}_{\boldsymbol{\ell}}) e^{i\boldsymbol{\omega}_{\boldsymbol{\ell}}'\mathbf{h}} = R_{\mathbf{m}}(\mathbf{h}) \end{aligned}$$

where  $\boldsymbol{\omega}_{\boldsymbol{\ell}} = (2\pi\ell_1/m_1, \dots, 2\pi\ell_d/m_d)$ . □

## References

- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Crc Press.
- Cressie, N. (1993). *Statistics for spatial data: Wiley series in probability and statistics*.
- Dietrich, C. and Newsam, G. N. (1997). Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, 18(4):1088–1107.
- Gneiting, T., Ševčíková, H., Percival, D. B., Schlather, M., and Jiang, Y. (2006). Fast and exact simulation of large Gaussian lattice systems in  $r_2$ : Exploring the limits. *Journal of Computational and Graphical Statistics*, 15(3).
- Guinness, J., Fuentes, M., Hesterberg, D., and Polizzotto, M. (2014). Multivariate spatial modeling of conditional dependence in microscale soil elemental composition data. *Spatial Statistics*, 9:93–108.
- 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.
- Matérn, B. (1960). *Spatial variation: Stochastic models and their application to some problems in forest surveys and other sampling investigations*. Meddelanden fran statens Skogsforskningsinstitut.
- Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC Press.
- Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer.
- Stein, M. L. (2002). Fast and exact simulation of fractional Brownian surfaces. *Journal of Computational and Graphical Statistics*, 11(3):587–599.
- 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.
- Stroud, J. R., Stein, M. L., and Lysen, S. (2014). Bayesian and maximum likelihood estimation for Gaussian processes on an incomplete lattice. *arXiv preprint arXiv:1402.4281*.
- Varin, C., Reid, N. M., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, pages 434–449.

Wood, A. T. A. and Chan, G. (1994). Simulation of stationary Gaussian processes in  $[0, 1]^d$ . *Journal of Computational and Graphical Statistics*, 3(4):409–432.