A Multi-Resolution Approximation for Massive Spatial Datasets

Matthias Katzfuss

Taylor & Francis
Taylor & Francis Group

A Multi-Resolution Approximation for Massive Spatial Datasets

Matthias Katzfuss

Department of Statistics, Texas A&M University, College Station, TX

ABSTRACT

Automated sensing instruments on satellites and aircraft have enabled the collection of massive amounts of high-resolution observations of spatial fields over large spatial regions. If these datasets can be efficiently exploited, they can provide new insights on a wide variety of issues. However, traditional spatial-statistical techniques such as kriging are not computationally feasible for big datasets. We propose a multi-resolution approximation (*M*-RA) of Gaussian processes observed at irregular locations in space. The *M*-RA process is specified as a linear combination of basis functions at multiple levels of spatial resolution, which can capture spatial structure from very fine to very large scales. The basis functions are automatically chosen to approximate a given covariance function, which can be nonstationary. All computations involving the *M*-RA, including parameter inference and prediction, are highly scalable for massive datasets. Crucially, the inference algorithms can also be parallelized to take full advantage of large distributed-memory computing environments. In comparisons using simulated data and a large satellite dataset, the *M*-RA outperforms a related state-of-the-art method. Supplementary materials for this article are available online.

1. Introduction

Automated sensing instruments on satellites and aircraft have enabled the collection of massive amounts of high-resolution observations of spatial fields over large and inhomogenous spatial domains. If these kinds of datasets can be efficiently exploited, they can provide new insights on a wide variety of issues, such as greenhouse gas concentrations for climate change, soil properties for precision agriculture, and atmospheric states for weather forecasting. Based (implicitly or explicitly) on Gaussian processes, the field of spatial statistics provides a rich toolkit for the analysis of such data, including estimating unknown parameters, predicting the spatial field at unobserved locations, and properly quantifying uncertainty in the predictions and parameters (e.g., Cressie and Wikle 2011).

However, traditional spatial-statistical techniques such as kriging are not computationally feasible for big datasets, because dense $n \times n$ matrices need to be decomposed, where n is the number of measurements. General-purpose methods such as the preconditioned conjugate gradient algorithm (e.g., Golub and Van Loan 2012) or probabilistic projections (e.g., Halko, Martinsson, and Tropp 2011; Stein, Chen, and Anitescu 2013) can solve large linear systems. But it is challenging for such algorithms to calculate the determinant of the data covariance matrix also required for likelihood-based inference, as this covariance is huge, often dense, and might have a slowly decaying spectrum.

More specialized methods for approximating spatial inference that explicitly try to exploit spatial information in the data have been proposed in the literature, but most of them either require restrictive assumptions about the covariance function

(e.g., Fuentes 2007; Lindgren, Rue, and Lindström 2011), or they ignore much of the fine-scale dependence (e.g., Higdon 1998; Mardia et al. 1998; Calder 2007; Cressie and Johannesson 2008; Katzfuss and Cressie 2009; Lemos and Sansó 2009; Katzfuss and Cressie 2011, 2012) or the large-scale dependence (e.g., Furrer, Genton, and Nychka 2006; Kaufman, Schervish, and Nychka 2008; Shaby and Ruppert 2012). Composite-likelihood methods (e.g., Vecchia 1988; Curriero and Lele 1999; Stein, Chi, and Welty 2004; Caragea and Smith 2007; Bevilacqua et al. 2012; Eidsvik et al. 2014) achieve computational feasibility by treating (blocks of) observations as conditionally independent, but it is not clear how to obtain proper joint predictive distributions for locations in different blocks.

We propose here a multi-resolution approximation (*M*-RA) of Gaussian processes observed at irregular (i.e., nongridded) locations in space. The *M*-RA process is specified as a linear combination of basis functions at multiple levels of spatial resolution, which can capture spatial structure from very fine to very large scales. Multi-resolution models (e.g., Chui 1992; Johannesson, Cressie, and Huang 2007; Nychka et al. 2016) have been very successful in spatial statistics, due to their ability to flexibly capture dependence at multiple spatial scales while being computationally feasible. In contrast to these existing methods, in our *M*-RA the basis functions and the distributions of their weights are chosen to "optimally" approximate a given covariance function, without requiring any restrictions on this covariance function. The basis functions in each region at each resolution are chosen iteratively according to the rules of the predictive process (Quiñonero Candela and Rasmussen 2005; Banerjee et al. 2008), based on a recursive partitioning of the

spatial domain into smaller and smaller subregions, and a set of "knot" locations in each region. The *M*-RA is a generalization of the full-scale approximation (Snelson and Ghahramani 2007; Sang, Jun, and Huang 2011; Sang and Huang 2012), a current state-of-the-art method for covariance approximations for large spatial data, which has only one level of domain partitioning and one resolution of basis functions. We will compare the two approaches extensively.

Inference for basis-function models essentially consists of obtaining the posterior distribution of the basis-function weights. In previous approaches, computationally feasible inference has been achieved by limiting the number of basis functions to be small (e.g., Higdon 1998; Quiñonero Candela and Rasmussen 2005; Cressie and Johannesson 2008) or by specifying the precision matrix of the weight vector to be diagonal or sparse (e.g., Higdon 1998; Lindgren, Rue, and Lindström 2011; Nychka et al. 2016). The *M*-RA combines both approaches: It results in a multi-resolution (block) sparse precision matrix, but the number of spatial basis functions within each region is small, allowing repeated application of the Sherman–Morrison–Woodbury formula (Sherman and Morrison 1950; Woodbury 1950). This leads to a highly scalable inference algorithm for the *M*-RA (which could also be applied to any multi-resolution basis-function model with a similar structure). Crucially, based on previous work (Katzfuss and Hammerling 2017) describing parallel algorithms for a special case of the *M*-RA, we derive algorithms that can split the computing task efficiently between many nodes. This way, spatial inference could be carried out for massive spatial datasets, using parallel computations at a number of nodes, each dealing only with a subset of the dataset. If enough computing nodes are available, this ensures scalability of the *M*-RA even for datasets with many millions of observations.

This article is organized as follows. In Section 2, we introduce the *M*-RA and discuss some of its properties. In Section 3, we present algorithms for parameter inference and spatial prediction, and describe the computational complexity of the *M*-RA. In Section 4, we apply the *M*-RA to large simulated datasets and to a real-data example and compare the *M*-RA to the full-scale approximation. We conclude in Section 5. All proofs are given in the Appendix.

2. Multi-Resolution Approximation (*M*-RA)

We begin this section by describing the true Gaussian process to be approximated (Section 2.1). After some preliminaries (Section 2.2), we introduce the multi-resolution approximation (*M*-RA) (Section 2.3) and discuss its properties (Sections 2.4 and 2.5).

2.1. The True Gaussian Process

Let $\{y_0(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$, or $y_0(\cdot)$, be the true spatial field or process of interest on a continuous (nongridded) domain $\mathcal{D} \subset \mathbb{R}^d, d \in \mathbb{N}^+$. We assume that $y_0(\cdot) \sim \mathrm{GP}(0, C_0)$ is a zero-mean Gaussian process with covariance function C_0. We place no restrictions on C_0, other than assuming that it is a positive-definite function on $\mathcal{D}$ that is known up to a vector of parameters, $\boldsymbol{\theta}$. In real applications, $y_0(\cdot)$ will often not have mean zero, but estimating

and subtracting the mean is usually not a computational problem. Once $y_0(\cdot)$ has been observed at a set of n spatial locations, $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, the distribution of the data is given by

$$\mathbf{y}_0(\mathcal{S}) := \big(y_0(\mathbf{s}_1), \ldots, y_0(\mathbf{s}_n)\big)' \sim N_n\big(\mathbf{0}, \mathbf{C}_0(\mathcal{S}, \mathcal{S})\big),$$

an n-variate Gaussian distribution with covariance matrix $\mathbf{C}_0(\mathcal{S}, \mathcal{S}) = \big(C_0(\mathbf{s}_i, \mathbf{s}_j)\big)_{i,j=1,\ldots,n}$.

The basic goal of spatial statistics is to make (likelihood-based) inference on the parameters $\boldsymbol{\theta}$ and to obtain spatial predictions of $y_0(\cdot)$ at a set of locations $\mathcal{S}^P$ (i.e., to obtain the posterior distribution of $\mathbf{y}_0(\mathcal{S}^P)$). This requires multiple Cholesky decompositions of the data covariance matrix $\mathbf{C}_0(\mathcal{S}, \mathcal{S})$, which generally has $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ memory complexity. This is computationally infeasible when $n = 10^5$ or more. In addition, computations for inference are also very difficult to parallelize, as computations with a dense and large covariance matrix require substantial communication overhead. Thus, the computational challenges cannot be solved by brute-force use of high-performance computing systems, and approximations or simplifying assumptions are necessary.

2.2. Domain Partitioning and Knots

To define the *M*-RA, we need a recursive partitioning of the spatial domain $\mathcal{D}$, in which each of J regions, $\mathcal{D}_1, \ldots, \mathcal{D}_J$, is again divided into J smaller subregions, and so forth, up to level M:

$$\mathcal{D}_{j_1, \ldots, j_{m-1}} = \dot{\bigcup}_{j_m=1,\ldots,J} \mathcal{D}_{j_1, \ldots, j_m}, \quad j_1, \ldots, j_{m-1} = 1, \ldots, J;$$

$$m = 1, \ldots, M.$$

For a generic Gaussian process $x(\cdot) \sim \mathrm{GP}(0, C)$, we define $[x(\cdot)]_{[m]}$ to be a "block-independent" version of $x(\cdot)$ between regions at resolution m, that is, $[x(\cdot)]_{[m]} \sim \mathrm{GP}(0, [C]_{[m]})$, where $[C]_{[m]}(\mathbf{s}_1, \mathbf{s}_2) = C(\mathbf{s}_1, \mathbf{s}_2)$ if $\mathbf{s}_1, \mathbf{s}_2$ are in the same region $\mathcal{D}_{j_1, \ldots, j_m}$ at resolution m, and $[C]_{[m]}(\mathbf{s}_1, \mathbf{s}_2) = 0$ otherwise.

We also need a multi-resolutional set of knots, such that $\mathcal{Q}_{j_1, \ldots, j_m}$ is a set of r knots (with $r << n$) that all lie in subregion $\mathcal{D}_{j_1, \ldots, j_m}$. For ease of notation, we assume that the knots in each of the regions at resolution M are given by the observation locations in that region: $\mathcal{Q}_{j_1, \ldots, j_M} = \mathcal{S}_{j_1, \ldots, j_M}$. Further, we write $\mathcal{Q}^{(m)} = \{\mathcal{Q}_{j_1, \ldots, j_m} : j_1, \ldots, j_m = 1, \ldots, J\}$ for the set of all rJ^m knots at resolution m. For a one-dimensional toy example, the top row of Figure 1 shows partitions and basis functions for *M*-RAs with $M = 1$ and $M = 3$. The knots are the locations at which the basis functions attain their maximum.

Note that we have assumed the same number (J) of subregions in each partition and the same number (r) of knots in each subregion, but this is only for notational convenience and not necessary for the inference described later. Hereafter, we will assume the domain partitioning and knots to be fixed and known. Some further discussion of their choice is given in Section 2.5.

2.3. Definition of the Multi-Resolution Approximation (M-RA)

Recall the true Gaussian process $y_0(\cdot) \sim \mathrm{GP}(0, C_0)$ from Section 2.1. We assume temporarily that the covariance function

(a) Simulated data (red dots), and $J = 9$ partitions (red vertical lines) and $r = 6$ basis functions (blue) for a 1-RA

(b) Data (red dots), and $J = 3$ partitions (vertical lines) and $r = 2$ basis functions per region for a 3-RA

(c) True covariance and 1-RA up to resolution m

(d) True covariance and M-RA up to resolution m

(e) Posterior predictive means (solid lines) and pointwise 95% posterior prediction intervals (dashed lines) for $y(\cdot)$. The interval bounded by the gray vertical lines is "unobserved."
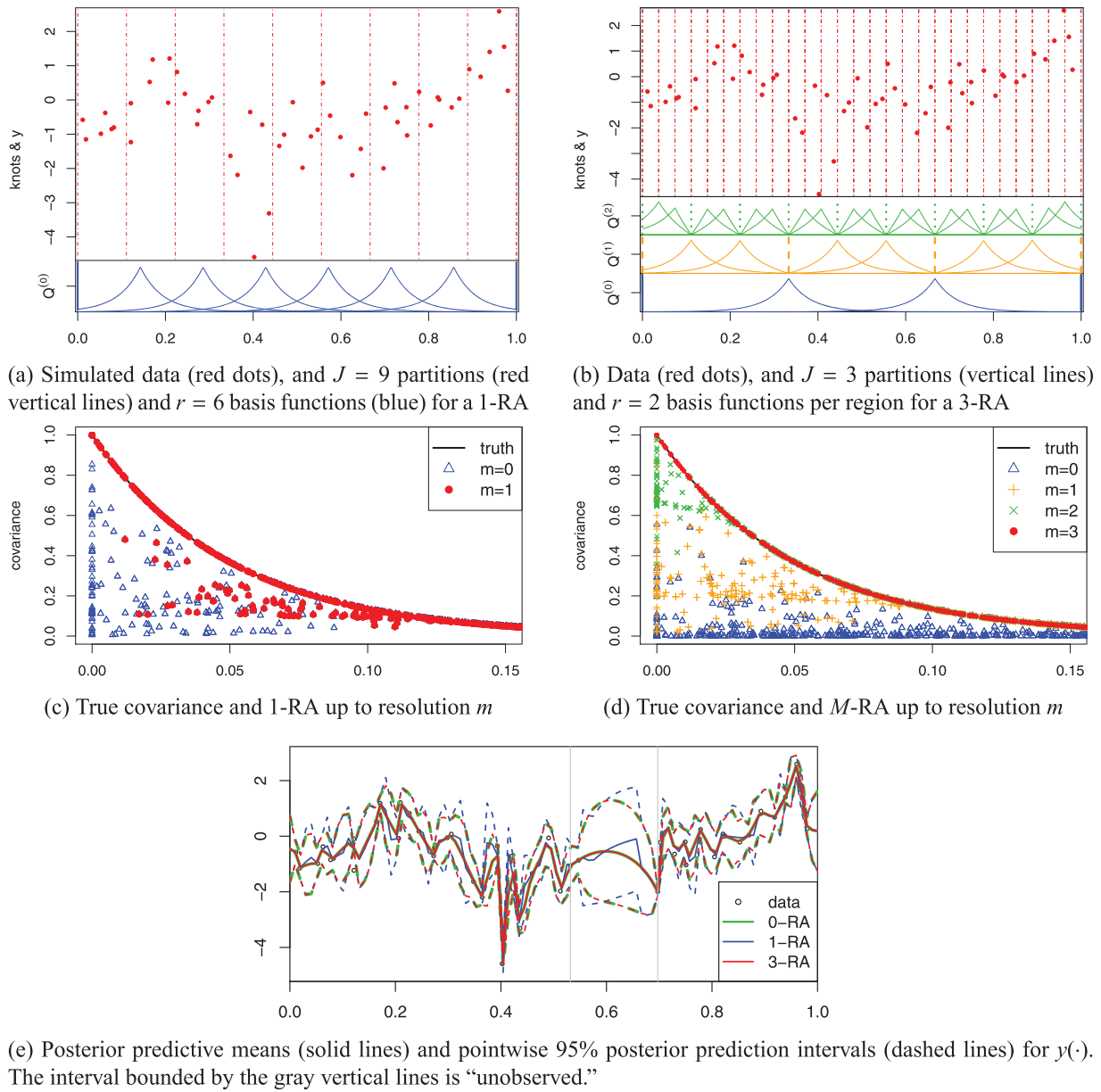
Figure 1. Comparison of a full-scale approximation (1-RA) to a multi-resolution approximation with three resolutions (3-RA) with the same computational complexity ($Mr = 6$ for both models) in a toy example of $n = 54$ observations generated from an exponential covariance function on a one-dimensional spatial domain $\mathcal{D} = [0, 1]$. Panels (c) and (d) show the covariance approximations using only the basis functions up to resolution m, for $m = 0, \ldots, M$. As m increases, deviations between the true and approximated covariance occur only on smaller and smaller scales (distances). For $m = M$, the covariance and predictions of the 3-RA are exactly the same as the true covariance and the corresponding (optimal) predictions of the 0-RA, whereas the covariance approximations and predictions using the 1-RA differ considerably from the truth. Note that for other covariance functions or in higher dimensions, the M-RA will not generally be exact (see Section 2.5).

C_0 is fully known (parameter inference is discussed in Section 3.3). The M-RA process approximates $y_0(\cdot)$ and its covariance C_0 iteratively at resolutions $m = 0, \ldots, M$, based on the knots and partitions specified in Section 2.2. At each resolution m, it approximates the "remainder" term—the difference between $y_0(\cdot)$ and the approximations at lower resolutions—using the predictive process (Banerjee et al. 2008), independently between regions $\mathcal{D}_{j_1, \ldots, j_m}$. As illustrated in Panel (d) of Figure 1, low M-RA resolutions capture variability at low frequencies (i.e., at large distances), resulting in remainder terms that exhibit variability on smaller and smaller scales as m increases, and so approximating the remainder independently between finer and finer partitions causes little approximation error.

More specifically, we begin with a predictive-process approximation of $y_0(\cdot)$, $\tau_0(\mathbf{s}) := E(y_0(\mathbf{s})|\mathbf{y}_0(\mathcal{Q}^{(0)}))$, $\mathbf{s} \in \mathcal{D}$, and we approximate the remainder process by assuming it to be independent between regions $\mathcal{D}_1, \ldots, \mathcal{D}_J$ at resolution 1: $\delta_1(\cdot) = [y_0(\cdot) - \tau_0(\cdot)]_{[1]}$ (Sang, Jun, and Huang 2011). We then again approximate this remainder process as the sum of its predictive process, $\tau_1(\mathbf{s}) = E(\delta_1(\mathbf{s})|\boldsymbol{\delta}_1(\mathcal{Q}^{(1)}))$, $\mathbf{s} \in \mathcal{D}$, plus the approximate remainder $\delta_2(\cdot) = [\delta_1(\cdot) - \tau_1(\cdot)]_{[2]}$, and so forth, up to level M. This leads to the following expression for the M-RA:

$$y_M(\cdot) = \tau_0(\cdot) + \tau_1(\cdot) + \cdots + \tau_{M-1}(\cdot) + \delta_M(\cdot), \qquad (1)$$

where $\tau_m(\mathbf{s}) = E(\delta_m(\mathbf{s})|\boldsymbol{\delta}_m(\mathcal{Q}^{(m)}))$, $\mathbf{s} \in \mathcal{D}$, and $\delta_m(\cdot) = [\delta_{m-1}(\cdot) - \tau_{m-1}(\cdot)]_{[m]} \sim \text{GP}(0, v_m)$.

An alternative expression for the M-RA in (1) can be obtained by noting that, for $m = 0, \ldots, M - 1$, we can write each predictive process for $\mathbf{s} \in \mathcal{D}_{j_1,\ldots,j_m}$ as a linear combination of basis functions (see Katzfuss 2013), $\tau_m(\mathbf{s}) = \mathbf{b}_{j_1,\ldots,j_m}(\mathbf{s})' \boldsymbol{\eta}_{j_1,\ldots,j_m}$ with $\boldsymbol{\eta}_{j_1,\ldots,j_m} \sim N_r(\mathbf{0}, \mathbf{K}_{j_1,\ldots,j_m})$, and

$$\mathbf{b}_{j_1,\ldots,j_m}(\mathbf{s}) := v_m(\mathbf{s}, \mathcal{Q}_{j_1,\ldots,j_m}), \quad \mathbf{s} \in \mathcal{D}_{j_1,\ldots,j_m}$$

$$\mathbf{K}_{j_1,\ldots,j_m}^{-1} := v_m(\mathcal{Q}_{j_1,\ldots,j_m}, \mathcal{Q}_{j_1,\ldots,j_m})$$

$$v_{m+1}(\mathbf{s}_1, \mathbf{s}_2) := v_m(\mathbf{s}_1, \mathbf{s}_2) - \mathbf{b}_{j_1,\ldots,j_m}(\mathbf{s}_1)' \mathbf{K}_{j_1,\ldots,j_m} \mathbf{b}_{j_1,\ldots,j_m}(\mathbf{s}_2),$$
$$\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}_{j_1,\ldots,j_m}, \quad (2)$$

where $v_{m+1}(\mathbf{s}_1, \mathbf{s}_2) = 0$ if $\mathbf{s}_1$ and $\mathbf{s}_2$ are in different regions at the mth resolution, and we set $v_0 = C_0$. Therefore, the M-RA can also be written as a linear combination of basis functions at M resolutions $0, \ldots, M - 1$, plus a remainder term at resolution M:

$$y_M(\mathbf{s}) = \mathbf{b}(\mathbf{s})' \boldsymbol{\eta} + \mathbf{b}_{j_1}(\mathbf{s})' \boldsymbol{\eta}_{j_1} + \cdots + \mathbf{b}_{j_1,\ldots,j_{M-1}}(\mathbf{s})' \boldsymbol{\eta}_{j_1,\ldots,j_{M-1}}$$
$$+ \delta_M(\mathbf{s}), \quad \mathbf{s} \in \mathcal{D}_{j_1,\ldots,j_M}, \quad (3)$$

where the weight vectors are independent of each other and of $\delta_M(\cdot) \sim \mathrm{GP}(0, v_M)$. Panels (a) and (c) of Figure 1 show the basis functions in the toy example. Once we have observed data at locations $\mathcal{S}$, we can also write the remainder $\delta_M(\cdot)$ in (3) as a linear combination of basis functions, $\delta_M(\mathbf{s}) = \mathbf{b}_{j_1,\ldots,j_M}(\mathbf{s})' \boldsymbol{\eta}_{j_1,\ldots,j_M}$, $\mathbf{s} \in \mathcal{D}_{j_1,\ldots,j_M}$, as in (2), where $\mathcal{Q}_{j_1,\ldots,j_M} = \mathcal{S}_{j_1,\ldots,j_M}$.

2.4. Properties of the M-RA

2.4.1. Many Basis Functions
In contrast to so-called low-rank approaches, which rely on a small or moderate number of basis functions and for which capturing small-scale variation is challenging (see, e.g., Stein 2014), the total number of basis functions for the M-RA with $M > 0$ is larger than the number of measurements: $r_{\text{total}} = r \sum_{m=0}^{M} J^M = rJ^M + r \sum_{m=0}^{M-1} J^m = n + r\frac{J^M - 1}{J - 1} > n$. This allows the M-RA to capture variation at all spatial scales, including very small scales.

2.4.2. Orthogonal Decomposition
Because the predictive process is a conditional expectation, which is a projection operator, the predictive process $\tau_m(\cdot)$ is independent of the remainder $\delta_m(\cdot) - \tau_m(\cdot)$, for all $m = 0, \ldots, M - 1$. Hence, we define the M-RA in (1) as a sum of orthogonal components. In the form (3), the M-RA is a weighted sum of spatial basis functions, for which the weights $\boldsymbol{\eta}_{j_1,\ldots,j_m}$ are block-orthogonal in probability space, but two sets of basis functions $\mathbf{b}_{j_1,\ldots,j_{m_1}}$ and $\mathbf{b}_{i_1,\ldots,i_{m_2}}$ are only block-orthogonal in physical space if $\mathcal{D}_{j_1,\ldots,j_{m_1}} \cap \mathcal{D}_{i_1,\ldots,i_{m_2}} = \varnothing$.

2.4.3. Valid Gaussian Process

Proposition 1. The M-RA is a valid Gaussian process with a non-negative definite covariance function, C_M.

2.4.4. "Optimal" Basis Functions
At every resolution $m = 0, \ldots, M - 1$ and within every region $\mathcal{D}_{j_1,\ldots,j_m}$, the goal of the M-RA in (1) is to approximate the remainder process $\delta_m(\cdot)$ as closely as possible, where

$$\delta_m(\cdot) = [\delta_{m-1}(\cdot) - \tau_{m-1}(\cdot)]_{[m]} = \Big[y_0(\cdot) - \sum_{l=0}^{m-1} \tau_l(\cdot)\Big]_{[m]}. \quad (4)$$

Hence, in each region, $\delta_m(\cdot)$ in (1) is the difference between the true process $y_0(\cdot)$ and the "previous" terms at lower resolutions, $\sum_{l=0}^{m-1} \tau_l(\cdot)$. We choose $\tau_m(\cdot)$ to be the predictive-process approximation of $\delta_m(\cdot)$. As this is a conditional expectation, $\tau_m(\cdot)$ is the function of $\delta_m(\mathcal{Q}^{(m)})$ that minimizes the expected squared difference to $\delta_m(\mathbf{s})$, conditional on $\delta_m(\mathcal{Q}^{(m)})$ (Banerjee et al. 2008). Further, $\tau_m(\cdot)$ can be viewed as an approximation of the optimal rank-r representation of $\delta_m(\cdot)$ within each region $\mathcal{D}_{j_1,\ldots,j_m}$, in that $\tau_m(\cdot)$ is the Nyström approximation of the first r terms in the Karhunen–Loève expansion of $\delta_m(\cdot)$ (Sang and Huang 2012). This is further evidence that at each resolution the predictive process captures variability at the low frequencies, leaving mostly higher-frequency variability to be captured at higher resolutions within smaller subregions. For increasing r, $\tau_m(\cdot)$ will be increasingly close to $\delta_m(\cdot)$. In fact, if $\mathcal{Q}_{j_1,\ldots,j_m}$ is equal to $\mathcal{S}_{j_1,\ldots,j_m}$, the set of observed locations in $\mathcal{D}_{j_1,\ldots,j_m}$, then it is straightforward to show that $\boldsymbol{\tau}_m(\mathcal{S}_{j_1,\ldots,j_m}) = \boldsymbol{\delta}_m(\mathcal{S}_{j_1,\ldots,j_m})$. In this sense, $\tau_m(\cdot)$ (and its basis-function representation) are an "optimal" approximation of $\delta_m(\cdot)$.

In contrast to many other multi-resolution methods for spatial data, the M-RA thus automatically provides a basis-function representation to approximate a given covariance function C_0 (based on a particular domain partitioning and set of knots), without any restrictions on C_0. This is illustrated in Figure 2, which shows the basis functions of a 3-RA for a highly nonstationary covariance function C_0.

2.4.5. Quality of the Covariance Approximation
At resolution m, the M-RA attempts to capture the covariance of the remainder $\delta_m(\cdot)$ between the partitions of each region $\mathcal{D}_{j_1,\ldots,j_m}$ by the predictive-process basis-functions component $\mathbf{b}_{j_1,\ldots,j_m}(\cdot)' \boldsymbol{\eta}_{j_1,\ldots,j_m}$. How close the covariance of the M-RA, $C_M(\mathbf{s}_1, \mathbf{s}_2)$, is to the true covariance, $C_0(\mathbf{s}_1, \mathbf{s}_2)$, depends on up to which resolution $\mathbf{s}_1$ and $\mathbf{s}_2$ lie in the same region. If $\mathbf{s}_1$ and $\mathbf{s}_2$ are in the same region at resolution M, then $C_M(\mathbf{s}_1, \mathbf{s}_2) = C_0(\mathbf{s}_1, \mathbf{s}_2)$. (To prove this, simply combine (1) with (4).) This also implies that the variances of $y_0(\cdot)$ and $y_M(\cdot)$ are the same. If $\mathbf{s}_1$ and $\mathbf{s}_2$ are in the same region $\mathcal{D}_{j_1,\ldots,j_m}$ at resolution $m < M$, but not at resolution $m + 1$, then $C_0(\mathbf{s}_1, \mathbf{s}_2)$ is only approximated by the basis functions up to resolution m: $C_M(\mathbf{s}_1, \mathbf{s}_2) = \sum_{l=0}^{m} \mathbf{b}_{j_1,\ldots,j_l}(\mathbf{s}_1)' \mathbf{K}_{j_1,\ldots,j_l} \mathbf{b}_{j_1,\ldots,j_l}(\mathbf{s}_2)$.

2.4.6. Comparison to the Full-Scale Approximation
Important special cases of the M-RA are the original process y_0 for $M = 0$, and the full-scale approximation (Snelson and Ghahramani 2007; Sang, Jun, and Huang 2011) for $M = 1$. The full-scale approximation, or 1-RA, only has basis functions at one resolution with r^F knots $\mathcal{Q}^F$, and a single level of domain

(a) True covariance matrix

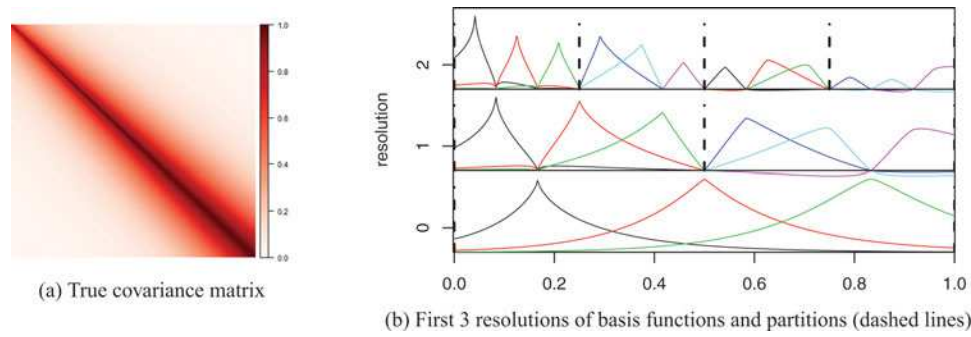(b) First 3 resolutions of basis functions and partitions (dashed lines)

Figure 2. For a spatial process with a nonstationary Matérn covariance function with range 0.15 and spatially varying (increasing) smoothness $\nu(s) = 0.2 + 0.7s$ on a one-dimensional domain $\mathcal{D} = [0, 1]$, basis functions, and partitions of an M-RA with $J = 2$ and $r = 3$. Note how the basis functions adapt to the increasing smoothness of the true covariance function and according to the placement of basis functions at lower resolutions.

partitioning, $\mathcal{D} = \dot{\bigcup}_{j=1,\dots,J^F} \mathcal{D}_j^F$. For massive datasets, the sub-regions $\mathcal{D}_j^F$ need to be very small to maintain computational feasibility.

Comparing our M-RA (with $M > 1$) and a full-scale approximation (1-RA) with $\mathcal{Q}^F = \mathcal{Q}$ and $\{\mathcal{D}_j^F : j = 1, \dots, J^F\} = \{\mathcal{D}_{j_1,\dots,j_M} : j_1, \dots, j_M = 1, \dots, J\}$, the covariance approximation for the two models is the same for pairs of locations in the same finest subregion $\mathcal{D}_j^F$ and for pairs in different subregions $\mathcal{D}_i$ and $\mathcal{D}_j$ with $i \neq j$ at the (coarsest) resolution 1. For all other pairs of locations, the M-RA has extra basis functions to capture their dependence, and if r is sufficiently large that $\tau_m(\cdot)$ captures the dependence of $\delta_m(\cdot)$ between subregions $\mathcal{D}_{j_1,\dots,j_{m+1}}$ well, the M-RA will provide a better approximation of $y_0(\cdot)$ than the 1-RA.

As described later in Section 3.6, the M-RA with r knots has the same computational complexity as the 1-RA with Mr knots. As illustrated in Figure 1, the M-RA can result in considerably better approximations. Further comparisons are presented in Section 4.

2.5. More on the Choice of Knots and Partitions

To achieve good approximations, we recommend choosing M and J as small and r as large as the computational resources allow (see Table 1(a)), under the constraint that $rJ^M \geq n$.

If the observation locations are approximately uniformly distributed over the domain $\mathcal{D}$, the partitions can simply be obtained by recursively splitting each region into J subregions of equal area. If the observation locations are far from uniform, more complicated partitioning schemes might be necessary to achieve fast inference.

Table 1. Time and memory complexity of the M-RA and its special cases, regular kriging (0-RA) and full-scale approximation (1-RA), on a single computer and in a distributed computing environment with many nodes. See Sections 3.5–3.6 for more details.

	Single processor		Distributed	
	Time	Memory	Time	Memory
(a) Increasing r and M				
0-RA	n^3	n^2		
1-RA	nr^2	nr	r^3	r^2
M-RA	$n(Mr)^2$	nMr	$(Mr)^3$	Mr^2
(b) $M = \mathcal{O}(\log n)$				
M-RA	$n\log^2 n$	$n\log n$	$\log^3 n$	$\log n$

The remaining issue is the placement of the r knots within each region. A simple solution is to use equidistant grids over each region $\mathcal{D}_{j_1,\dots,j_m}$, but it can also be advantageous to place more knots close to the boundaries within each region. To see why, remember from (4) and Section 2.3 that the goal within each region $\mathcal{D}_{j_1,\dots,j_m}$ is to approximate $\delta_m(\cdot) \sim \text{GP}(0, v_m)$. Consider the case of a region with $J = 2$ subregions containing observed locations $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2\}$ with $\mathcal{S}_j = \{\mathcal{S}_j^B, \mathcal{S}_j^I\}$, where $\mathcal{S}_j^B$ are the locations within a distance c from the boundary and $\mathcal{S}_j^I$ are the remaining locations in the interior of subregion j. Choosing the knots $\mathcal{Q}_{j_1,\dots,j_m} = \mathcal{S}^B := \{\mathcal{S}_1^B, \mathcal{S}_2^B\}$, it can be shown that $\text{var}(\delta_m(\mathcal{S})) = \text{var}(\tau_m(\mathcal{S})) + \text{var}(\delta_m(\mathcal{S}) | \delta_m(\mathcal{S}^B))$, the latter being a matrix with only one nonzero block, $\text{var}(\delta_m(\mathcal{S}^I) | \delta_m(\mathcal{S}^B))$. The only part of this matrix that is ignored by the M-RA is $\text{cov}(\delta_m(\mathcal{S}_1^I), \delta_m(\mathcal{S}_2^I) | \delta_m(\mathcal{S}^B))$, which should be very small if c is large and/or the screening effect (e.g., Stein 2011) holds for v_m.

An extreme case of this strategy is illustrated in Figure 1. For the exponential covariance function without nugget in one spatial dimension, the screening effect holds exactly, in that two observations are conditionally independent given a third observation that separates the two. Because the knots for a particular resolution in Figure 1 are placed on the boundaries between partitions at the next higher resolution, the M-RA is exact in this case. For covariance functions without screening effect or in higher dimensions, the M-RA will generally not be exact.

While the (favorable) numerical results in Section 4.1 are obtained with the simplest choice of equal-area partitions and equally spaced knots, it is possible to adopt more complicated strategies, such as choosing the knots and partitions based on clustering (e.g., Snelson and Ghahramani 2007) or using reversible-jump Markov chain Monte Carlo (e.g., Gramacy and Lee 2008; Katzfuss 2013). Any potential boundary effects due to the choice of partitions can be alleviated by carrying out several M-RAs with different, shifted partitions and combining the results using Bayesian model averaging (e.g., Hoeting et al. 1999).

3. Inference

In this section, we describe inference for the M-RA. For a particular value of the parameter vector $\boldsymbol{\theta}$, the covariance function C_0, and hence the basis functions $\mathbf{b}_{j_1,\dots,j_m}$ and the covariance matrices $\mathbf{K}_{j_1,\dots,j_m}$ in (3) are fixed. The prerequisite for inference

is to calculate the quantities summarizing the prior distribution induced by the M-RA at the chosen knots and observed locations (Section 3.1). Then, the main task for inference is to obtain the posterior distribution of the unknown weight vectors $\mathcal{E}_{M-1}$ (Section 3.2), where we define $\mathcal{E}_m := \{\boldsymbol{\eta}_{j_1,\ldots,j_l} : j_1,\ldots,j_l = 1,\ldots,J; \, l = 0,\ldots,m\}$ for all $m = 0,\ldots,M-1$ to be the set of all basis-function weights at resolution m and all lower resolutions (and we let $\mathcal{E}_{-1} = \oslash$ be the empty set). Once this posterior distribution has been obtained, it can be used to evaluate the likelihood (Section 3.3) and to obtain spatial predictions (Section 3.4). By exploiting the block-sparse multi-resolution structure of the prior and posterior precision matrices of the weights, we obtain inference that has excellent time and memory complexity (Section 3.6), can take full advantage of distributed-memory systems with a large number of nodes (Section 3.5), and is thus scalable to massive spatial datasets.

3.1. Calculating the Prior Quantities

Let $\mathcal{S}_{j_1,\ldots,j_m}$ be the observation locations that lie in region $\mathcal{D}_{j_1,\ldots,j_m}$, and define

$$\mathbf{B}^l_{j_1,\ldots,j_m} := \mathbf{b}_{j_1,\ldots,j_l}(\mathcal{S}_{j_1,\ldots,j_m}), \quad l = 0,1,\ldots,m,$$

$$\boldsymbol{\Sigma}_{j_1,\ldots,j_m} := \mathrm{var}(\mathbf{y}_M(\mathcal{S}_{j_1,\ldots,j_m})|\mathcal{E}_{m-1})$$

$$= \mathbf{B}^m_{j_1,\ldots,j_m}\mathbf{K}_{j_1,\ldots,j_m}\mathbf{B}^m_{j_1,\ldots,j_m}{}' + \mathbf{V}_{j_1,\ldots,j_m},$$

$$\mathbf{V}_{j_1,\ldots,j_m} := \mathrm{var}(\mathbf{y}_M(\mathcal{S}_{j_1,\ldots,j_m})|\mathcal{E}_m)$$

$$= \mathrm{blockdiag}\left\{\boldsymbol{\Sigma}_{j_1,\ldots,j_m,1},\ldots,\boldsymbol{\Sigma}_{j_1,\ldots,j_m,J}\right\}, \qquad (5)$$

for $m = 0,1,\ldots,M-1$, and $\boldsymbol{\Sigma}_{j_1,\ldots,j_M} := v_M(\mathcal{S}_{j_1,\ldots,j_M}, \mathcal{S}_{j_1,\ldots,j_M})$.

For inference, we explicitly need to calculate the matrices $\{\mathbf{K}^{-1}_{j_1,\ldots,j_m} : j_1,\ldots,j_m = 1,\ldots,J; \, m = 0,\ldots,M-1\}$, $\{\mathbf{B}^l_{j_1,\ldots,j_M} : j_1,\ldots,j_M = 1,\ldots,J; l = 0,\ldots,M-1\}$, and $\{\boldsymbol{\Sigma}_{j_1,\ldots,j_M} : j_1,\ldots,j_M = 1,\ldots,J\}$. Defining $\mathbf{W}^l_{j_1,\ldots,j_m} := v_l(\mathcal{Q}_{j_1,\ldots,j_m}, \mathcal{Q}_{j_1,\ldots,j_l})$, we can do so by calculating

$$\mathbf{W}^l_{j_1,\ldots,j_m} = C_0(\mathcal{Q}_{j_1,\ldots,j_m}, \mathcal{Q}_{j_1,\ldots,j_l})$$

$$- \sum_{k=0}^{l-1} \mathbf{W}^k_{j_1,\ldots,j_m} \mathbf{K}_{j_1,\ldots,j_k} \mathbf{W}^k_{j_1,\ldots,j_l}{}' \qquad (6)$$

for $m = 0,\ldots,M$, $j_1,\ldots,j_m = 1,\ldots,J$, and $l = 0,\ldots,m$. Then we have $\mathbf{K}^{-1}_{j_1,\ldots,j_m} = \mathbf{W}^m_{j_1,\ldots,j_m}$ for $m < M$, $\mathbf{B}^l_{j_1,\ldots,j_M} = \mathbf{W}^l_{j_1,\ldots,j_M}$ for $l < M$, and $\boldsymbol{\Sigma}_{j_1,\ldots,j_M} = \mathbf{W}^M_{j_1,\ldots,j_M}$.

As an aside, other parameterizations of these matrices (and the quantities in (3)) are also possible and will lead to similar inference algorithms as described later, as long as the weight vectors are a priori independent and the basis functions have the same limited support.

3.2. The Posterior Distribution of the Basis-Function Weights

The definition of the M-RA in (3), together with the definitions in (5), implies that

$$\mathbf{y}_M\left(\mathcal{S}_{j_1,\ldots,j_m}\right)|\mathcal{E}_m \sim N\left(\sum_{l=0}^m \mathbf{B}^l_{j_1,\ldots,j_m}\boldsymbol{\eta}_{j_1,\ldots,j_l}, \mathbf{V}_{j_1,\ldots,j_m}\right).$$

Using the results in Katzfuss and Hammerling (2017, sec. 3), it can be shown that the conditional posterior distributions of the weight vectors for all $m = 0,\ldots,M-1$ are given by

$$\boldsymbol{\eta}_{j_1,\ldots,j_m}|\mathbf{y}_M(\mathcal{S}),\mathcal{E}_{m-1} \overset{\text{ind.}}{\sim} N_r\left(\widetilde{\boldsymbol{v}}_{j_1,\ldots,j_m}, \widetilde{\mathbf{K}}_{j_1,\ldots,j_m}\right),$$

$$j_1,\ldots,j_m = 1,\ldots,J, \qquad (7)$$

with posterior precision and mean

$$\widetilde{\mathbf{K}}^{-1}_{j_1,\ldots,j_m} = \mathbf{K}^{-1}_{j_1,\ldots,j_m} + \mathbf{B}^m_{j_1,\ldots,j_m}{}'\mathbf{V}^{-1}_{j_1,\ldots,j_m}\mathbf{B}^m_{j_1,\ldots,j_m}$$

$$= \mathbf{K}^{-1}_{j_1,\ldots,j_m} + \mathbf{A}^{m,m}_{j_1,\ldots,j_m},$$

$$\widetilde{\boldsymbol{v}}_{j_1,\ldots,j_m} = \widetilde{\mathbf{K}}_{j_1,\ldots,j_m}\left(\mathbf{B}^m_{j_1,\ldots,j_m}{}'\mathbf{V}^{-1}_{j_1,\ldots,j_m}(\mathbf{y}_M(\mathcal{S}_{j_1,\ldots,j_m})\right.$$

$$\left. - \sum_{l=0}^{m-1}\mathbf{B}^l_{j_1,\ldots,j_m}\boldsymbol{\eta}_{j_1,\ldots,j_l})\right)$$

$$= \widetilde{\mathbf{K}}_{j_1,\ldots,j_m}\boldsymbol{\omega}^m_{j_1,\ldots,j_m} - \sum_{l=0}^{m-1}\mathbf{K}_{j_1,\ldots,j_m}\mathbf{A}^{m,l}_{j_1,\ldots,j_m}\boldsymbol{\eta}_{j_1,\ldots,j_l}$$

$$(8)$$

respectively, where

$$\mathbf{A}^{k,l}_{j_1,\ldots,j_m} := \mathbf{B}^k_{j_1,\ldots,j_m}{}'\mathbf{V}^{-1}_{j_1,\ldots,j_m}\mathbf{B}^l_{j_1,\ldots,j_m} = \sum_{j_{m+1}=1}^J \widetilde{\mathbf{A}}^{k,l}_{j_1,\ldots,j_{m+1}},$$

$$k \geq l = 0,\ldots,m,$$

$$\boldsymbol{\omega}^k_{j_1,\ldots,j_m} := \mathbf{B}^k_{j_1,\ldots,j_m}{}'\mathbf{V}^{-1}_{j_1,\ldots,j_m}\mathbf{y}_M(\mathcal{S}_{j_1,\ldots,j_m}) = \sum_{j_{m+1}=1}^J \widetilde{\boldsymbol{\omega}}^k_{j_1,\ldots,j_{m+1}},$$

$$(9)$$

can be obtained recursively for $m = M-1, M-2,\ldots,0$ using

$$\widetilde{\mathbf{A}}^{k,l}_{j_1,\ldots,j_m} := \mathbf{B}^k_{j_1,\ldots,j_m}{}'\boldsymbol{\Sigma}^{-1}_{j_1,\ldots,j_m}\mathbf{B}^l_{j_1,\ldots,j_m}$$

$$= \mathbf{A}^{k,l}_{j_1,\ldots,j_m} - \mathbf{A}^{k,m}_{j_1,\ldots,j_m}\widetilde{\mathbf{K}}_{j_1,\ldots,j_m}\mathbf{A}^{m,l}_{j_1,\ldots,j_m},$$

$$\widetilde{\boldsymbol{\omega}}^k_{j_1,\ldots,j_m} := \mathbf{B}^k_{j_1,\ldots,j_m}{}'\boldsymbol{\Sigma}^{-1}_{j_1,\ldots,j_m}\mathbf{y}(\mathcal{S}_{j_1,\ldots,j_m})$$

$$= \boldsymbol{\omega}^k_{j_1,\ldots,j_m} - \mathbf{A}^{k,m}_{j_1,\ldots,j_m}\widetilde{\mathbf{K}}_{j_1,\ldots,j_m}\boldsymbol{\omega}^m_{j_1,\ldots,j_m}. \qquad (10)$$

In practice, the quantities in (10) are calculated directly from the definition (first equality) for $m = M$, and using the recursive expression (second equality) for $m = M-1,\ldots,0$. The proof of results (9)–(10) is straightforward using the Sherman–Morrison–Woodbury formula (Sherman and Morrison 1950; Woodbury 1950; Henderson and Searle 1981):

$$\boldsymbol{\Sigma}^{-1}_{j_1,\ldots,j_m} = \mathbf{V}^{-1}_{j_1,\ldots,j_m} - \mathbf{V}^{-1}_{j_1,\ldots,j_m}\mathbf{B}^m_{j_1,\ldots,j_m}\widetilde{\mathbf{K}}_{j_1,\ldots,j_m}\mathbf{B}^m_{j_1,\ldots,j_m}{}'\mathbf{V}^{-1}_{j_1,\ldots,j_m}.$$

$$(11)$$

3.3. Parameter Inference

Inference for the M-RA is based on $L_M(\boldsymbol{\theta})$, the likelihood of the observations $\mathbf{y}_M(\mathcal{S}) \sim N_n(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the M-RA covariance matrix given in (5) with $m = 0$ based on parameter vector $\boldsymbol{\theta}$:

$$-2 \log L_M(\boldsymbol{\theta}) = \log |\boldsymbol{\Sigma}| + \mathbf{y}_M(\mathcal{S})' \, \boldsymbol{\Sigma}^{-1} \, \mathbf{y}_M(\mathcal{S}).$$

This likelihood can be calculated using the quantities in Section 3.2. We have $-2 \log L_M(\boldsymbol{\theta}) = d + u$, with

$$d_{j_1,\dots,j_m} := \log |\boldsymbol{\Sigma}_{j_1,\dots,j_m}|,$$
$$u_{j_1,\dots,j_m} := \mathbf{y}_M(\mathcal{S}_{j_1,\dots,j_m})' \boldsymbol{\Sigma}_{j_1,\dots,j_m}^{-1} \mathbf{y}_M(\mathcal{S}_{j_1,\dots,j_m}),$$

$m = 0, 1, \dots, M$. For $m = M$, these quantities are calculated using the definition. For $m = M - 1, \dots, 0$, the recursive expressions

$$d_{j_1,\dots,j_m} = \log |\widetilde{\mathbf{K}}_{j_1,\dots,j_m}^{-1}| - \log |\mathbf{K}_{j_1,\dots,j_m}^{-1}| + \sum_{j_{m+1}=1}^{J} d_{j_1,\dots,j_{m+1}},$$
$$u_{j_1,\dots,j_m} = -\boldsymbol{\omega}_{j_1,\dots,j_m}^{m}{}' \widetilde{\mathbf{K}}_{j_1,\dots,j_m} \boldsymbol{\omega}_{j_1,\dots,j_m}^{m} + \sum_{j_{m+1}=1}^{J} u_{j_1,\dots,j_{m+1}},$$

can be derived using a matrix determinant lemma (e.g., Harville 1997, Thm. 18.1.1) and the Sherman–Morrison–Woodbury formula in (11).

In summary, the M-RA log-likelihood can be written as a sum of log-determinants and quadratic forms involving only $r \times r$ matrices. This result enables fast and scalable evaluation of the likelihood, which in turn allows for a wide array of likelihood-based inference techniques for an unknown parameter vector $\boldsymbol{\theta}$, such as maximum likelihood estimation, Markov chain Monte Carlo (MCMC), or particle filtering in spatio-temporal contexts.

3.4. Spatial Prediction

Spatial prediction can be carried out separately, after parameter inference is completed. In a frequentist context, prediction only has to be carried out once, for the final parameter estimates. In a Bayesian framework, parameter inference can be carried out only for the thinned MCMC chain, or for particles with considerable weight in the case of a particle sampler.

Implicitly conditioning on a particular value of the parameter vector $\boldsymbol{\theta}$, spatial prediction amounts to finding the posterior predictive distribution, $\mathbf{y}_M(\mathcal{S}^P)|\mathbf{y}_M(\mathcal{S})$, at a set of prediction locations $\mathcal{S}^P$. We denote by $\mathcal{S}_{j_1,\dots,j_M}^{P}$ the prediction locations in region $\mathcal{D}_{j_1,\dots,j_M}$. As a first step, we need to calculate prior prediction quantities similar to (6),

$$\mathbf{U}_{j_1,\dots,j_M}^{l} := v_l \left(\mathcal{S}_{j_1,\dots,j_M}^{P}, \mathcal{Q}_{j_1,\dots,j_l} \right)$$
$$= C_0 \left(\mathcal{S}_{j_1,\dots,j_M}^{P}, \mathcal{Q}_{j_1,\dots,j_l} \right) - \sum_{k=0}^{l-1} \mathbf{U}_{j_1,\dots,j_M}^{k} \mathbf{K}_{j_1,\dots,j_k} \mathbf{W}_{j_1,\dots,j_l}^{k}{}',$$

for $l = 0, \dots, M$, and then we set $\mathbf{L}_{j_1,\dots,j_M}^{M} := v_M(\mathcal{S}_{j_1,\dots,j_M}^{P}, \mathcal{S}_{j_1,\dots,j_M}) = \mathbf{U}_{j_1,\dots,j_M}^{M}$ and

$$\mathbf{V}_{j_1,\dots,j_M}^{P} := v_M \left(\mathcal{S}_{j_1,\dots,j_M}^{P}, \mathcal{S}_{j_1,\dots,j_M}^{P} \right)$$
$$= C_0 \left(\mathcal{S}_{j_1,\dots,j_M}^{P}, \mathcal{S}_{j_1,\dots,j_M}^{P} \right) - \sum_{k=0}^{M-1} \mathbf{U}_{j_1,\dots,j_M}^{k} \mathbf{K}_{j_1,\dots,j_k} \mathbf{U}_{j_1,\dots,j_M}^{k}{}'.$$

Spatial predictions can then be obtained using the following proposition.

Proposition 2. The posterior predictive distribution can be written as

$$\mathbf{y}_M \left(\mathcal{S}_{j_1,\dots,j_M}^{P} \right) | \mathbf{y}_M(\mathcal{S}) = \sum_{m=0}^{M-1} \widetilde{\mathbf{B}}_{j_1,\dots,j_M}^{m+1,m} \widetilde{\boldsymbol{\eta}}_{j_1,\dots,j_m} + \widetilde{\boldsymbol{\delta}}_{j_1,\dots,j_M}, \tag{12}$$

where

$$\widetilde{\boldsymbol{\eta}}_{j_1,\dots,j_m} \overset{\text{ind.}}{\sim} N_r \left(\widetilde{\mathbf{K}}_{j_1,\dots,j_m} \boldsymbol{\omega}_{j_1,\dots,j_m}^{m}, \widetilde{\mathbf{K}}_{j_1,\dots,j_m} \right),$$
$$\widetilde{\boldsymbol{\delta}}_{j_1,\dots,j_M} \overset{\text{ind.}}{\sim} N \left(\mathbf{L}_{j_1,\dots,j_M}^{M} \boldsymbol{\Sigma}_{j_1,\dots,j_M}^{-1} \mathbf{y}_M \left(\mathcal{S}_{j_1,\dots,j_M} \right), \right.$$
$$\left. \mathbf{V}_{j_1,\dots,j_M}^{P} - \mathbf{L}_{j_1,\dots,j_M}^{M} \boldsymbol{\Sigma}_{j_1,\dots,j_M}^{-1} \mathbf{L}_{j_1,\dots,j_M}^{M}{}' \right),$$

and the "posterior basis-function matrices" are given by

$$\widetilde{\mathbf{B}}_{j_1,\dots,j_M}^{l,k} := \mathbf{b}_{j_1,\dots,j_k} \left(\mathcal{S}_{j_1,\dots,j_M}^{P} \right) - \mathbf{L}_{j_1,\dots,j_l}^{l} \boldsymbol{\Sigma}_{j_1,\dots,j_l}^{-1} \mathbf{B}_{j_1,\dots,j_l}^{k}$$
$$= \widetilde{\mathbf{B}}_{j_1,\dots,j_M}^{l+1,k} - \widetilde{\mathbf{B}}_{j_1,\dots,j_M}^{l+1,l} \widetilde{\mathbf{K}}_{j_1,\dots,j_l} \mathbf{A}_{j_1,\dots,j_l}^{l,k}, \tag{13}$$

for $k = 0, 1, \dots, l - 1$.

Hence, the posterior predictive distribution in (12) has the same form as the (prior) M-RA process in (3). This allows calculation and storage of the *joint posterior predictive distribution*. Often, interest is in summaries of this joint posterior predictive distribution, such as the marginal posterior predictive distributions at each prediction location. In practice, the posterior basis-function matrices in (13) are calculated directly from the definition (first equation) for $l = M$, and using the recursive relation (second equation) for $l = M - 1, \dots, 0$.

3.5. Distributed Computing

A major advantage of the M-RA is that it is well suited to modern computing environments, in that computations can be carried out in a distributed fashion with little communication overhead at a large number of nodes, each only dealing with a small subset of the data.

More specifically, assume that we have nodes $\{\mathcal{N}_{j_1,\dots,j_m} : j_1, \dots, j_m = 1, \dots, J; \; m = 0, 1, \dots, M\}$ in a tree-like structure, as illustrated in Figure 3. Each node $\mathcal{N}_{j_1,\dots,j_m}$ holds the r knots or observation locations $\mathcal{Q}_{j_1,\dots,j_m}$ located in subregion $\mathcal{D}_{j_1,\dots,j_m}$, and it only has to work with matrices of size $r \times r$ (implying excellent load balance). The main computational effort for node $\mathcal{N}_{j_1,\dots,j_m}$ is in computing the Cholesky decomposition of the $r \times r$ matrix $\widetilde{\mathbf{K}}_{j_1,\dots,j_m}$ and calculating the quantities in (10), the latter of which could be parallelized if the node has multiple cores. The communication to each node $\mathcal{N}_{j_1,\dots,j_m}$ is $\mathcal{O}(Jm^2r^2)$, as it receives the matrices to calculate the quantities in (9) from its children. The calculations at the nodes/subregions for each resolution can be carried out completely in parallel.

For spatial prediction at locations $\mathcal{S}^P$, each (terminal) node $\mathcal{N}_{j_1,\dots,j_M}$ carries out parallel computations involving $\mathcal{S}_{j_1,\dots,j_M}^{P}$, the prediction locations in region $\mathcal{D}_{j_1,\dots,j_M}$, to obtain the "posterior basis-function matrices" in (13).
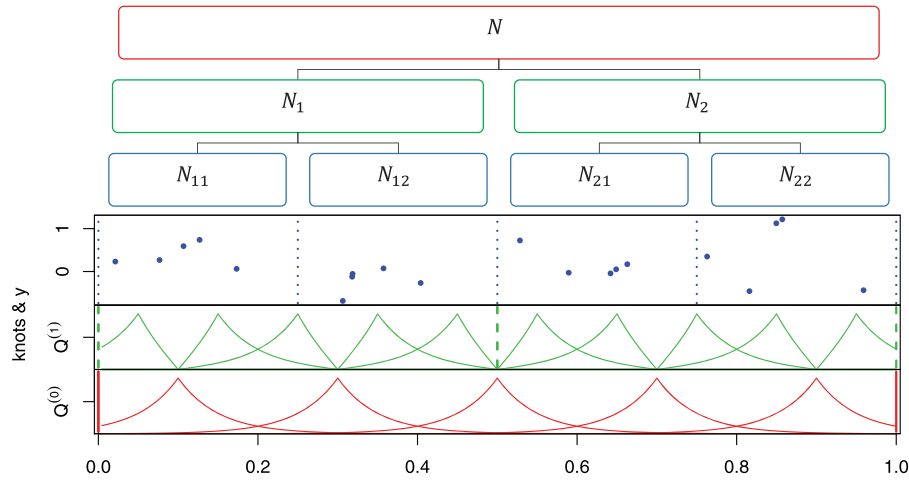
Figure 3. Illustration of the computational setup for distributed inference in the 2-RA with $J = 2$ and $r = 5$ for a toy example in a one-dimensional spatial domain. As indicated by matching colors, node $\mathcal{N}_{j_1,\ldots,j_m}$ works with the knots $\mathcal{Q}_{j_1,\ldots,j_m}$, $j_m = 1, 2$; $m = 0, 1, 2$. Only communication between connected nodes is necessary.

3.6. Computational Complexity

Remember that we assume here for simplicity that there is an equal number of $r = \frac{n}{J^M}$ knots or observation locations in each region $\mathcal{D}_{j_1,\ldots,j_m}$, and we regard J as a fixed (small) number. For each of the $\sum_{m=0}^{M} J^m < 2J^M$ regions, the main computational effort for inference is in obtaining the matrices $\widetilde{\mathbf{A}}_{j_1,\ldots,j_m}^{k,l}$ in (10), which requires one Cholesky decomposition of the $r \times r$ matrix $\widetilde{\mathbf{K}}_{j_1,\ldots,j_m}$ and computing $\mathcal{O}(m^2)$ quadratic forms of size $r \times r$. Thus, M-RA inference has $\mathcal{O}(J^M M^2 r^3) = \mathcal{O}(nr^2 M^2)$ time complexity and $\mathcal{O}(nrM)$ memory complexity.

When the M-RA is implemented in a distributed environment with a large number of nodes (as in Section 3.5 above), the overall time complexity is $\mathcal{O}(M^3 r^3)$ and the memory complexity per node is $\mathcal{O}(Mr^2)$, assuming that communication (which is $\mathcal{O}(M^2 r^2)$ per node) does not dominate computation time.

Thus, the M-RA with r knots has the same computational complexity as the 1-RA with Mr knots, but the M-RA can provide a much better approximation (see Figure 1). As is further explored in Section 4.1, as n is increasing, the performance of the 1-RA degrades unless r is allowed to increase as some fraction of n, while for the M-RA we can keep r fixed and instead let M increase with n as $M = \mathcal{O}(\log_J n)$. This is a natural assumption under increasing-domain asymptotics, for which an increase in the domain and data size by a factor of J allows an additional split of the resulting domain (i.e., increasing M by one) without degrading the approximation within the J subregions at the first resolution. In this case, the time and memory complexity for the M-RA are $\mathcal{O}(n \log^2 n)$ and $\mathcal{O}(n \log n)$, respectively, in the nondistributed setting. In the distributed setting, the overall time complexity is then $\mathcal{O}(\log^3 n)$, and the memory and communication complexity per node are $\mathcal{O}(\log n)$ and $\mathcal{O}(\log^2 n)$, respectively.

The same complexities hold for prediction, as long as the number of prediction locations per terminal region is on the same order as the number of observed locations (i.e., $|\mathcal{S}_{j_1,\ldots,j_M}^P| = \mathcal{O}(r)$). In addition, the M-RA allows us to store the entire joint predictive distribution in $\mathcal{O}(Mr^2 J^M) = \mathcal{O}(nMr)$ memory ($\mathcal{O}(Mr^2)$ per node in the distributed case). If $M = \mathcal{O}(\log n)$, this is $\mathcal{O}(n \log n)$ (or $\mathcal{O}(\log n)$ per node).

As summarized in Table 1(b), if we let M increase as $\log n$, the time and memory complexity for the M-RA are both quasi-linear in n, and even polylogarithmic in distributed settings with many nodes. Hence, the M-RA is highly scalable and can handle truly massive spatial datasets if enough computational nodes are available.

4. Numerical Comparisons and Illustrations

Using simulated and real data, we compared our proposed M-RA to the full-scale approximation (1-RA), which is a special case of the M-RA and a current state-of-the-art method for large spatial data. A nondistributed implementation of the methods in Julia (*http://julialang.org/*) version 0.3.7 was run on a 16-core machine (Intel Xeon 2.90GHz) with 64GB RAM. All Julia code, R code to produce the plots, and the data for Section 4.2 are available as supplementary material.

4.1. Simulation Study

We simulated five datasets, each roughly of size two million (specifically, $n_{\max} = 1{,}966{,}080$) from a Gaussian process with mean zero and covariance function

$$C_0(s_1, s_2) = 0.95 \, \mathcal{M}_{1.5}(|s_1 - s_2|/0.05) + 0.05 \, I(s_1 = s_2),$$

$$s_1, s_2 \in \mathcal{D},$$

where $I(\cdot)$ is the indicator function and

$$\mathcal{M}_{1.5}(h) = \left(1 + h\sqrt{3}\right) \exp\left(-h\sqrt{3}\right), \quad h \in \mathbb{R}_0^+ \quad (14)$$

is a Matérn correlation function with smoothness parameter 1.5, which is also used for the real-data example below in Section 4.2. The data were simulated on an equidistant grid on a one-dimensional domain $\mathcal{D} = [0, 1]$, which permitted fast simulation using the Davies–Harte algorithm and evaluation of the exact likelihood using the Durbin–Levinson algorithm for comparison. These algorithms were implemented in Julia along the lines of the functions `DHsimulate` and `DLLoglikelihood` in the R package `ltsa` (McLeod, Yu, and Krougly 2007).

From the "full" datasets of size $n_{\max}$, we created datasets of varying sample sizes n roughly between 2000 and 2 million.

(a) Loglikelihood under fixed domain

(b) Loglikelihood under increasing domain

(c) Times for likelihood evaluations
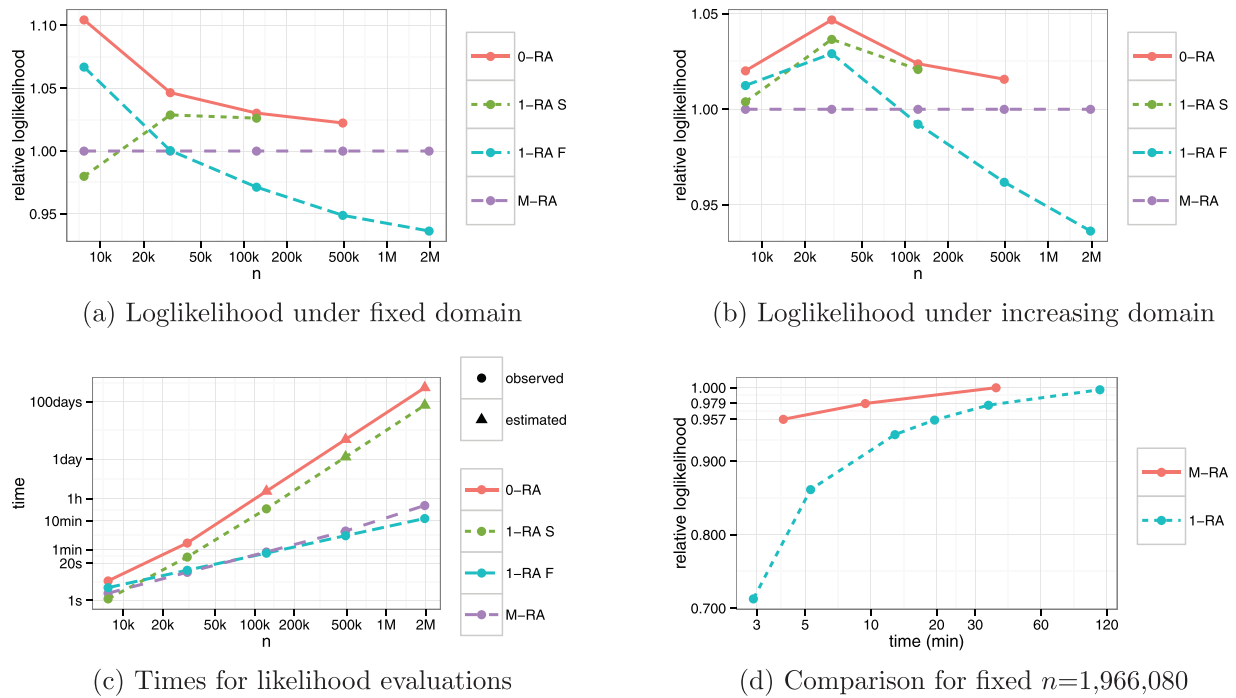

(d) Comparison for fixed $n=1{,}966{,}080$

Figure 4. Results from the simulation study in one spatial dimension described in Section 4.1. 0-RA is the true Gaussian process, and 1-RA S and 1-RA F are full-scale approximations with increasing and fixed r, respectively. Note that all axes indicating time or sample size are on a log scale. The log-scores (i.e., log-likelihoods) are all scaled relative to the log-score of the *M*-RA.

We considered the two types of asymptotics commonly used in spatial statistics: For fixed-domain (infill) asymptotics, we took equally spaced subsets of size n from the full dataset on the entire domain $[0, 1]$. For increasing-domain asymptotics, we always took the first n observations from the entire set.

We then recorded the log-likelihood and the time taken to compute it for each n and for each of the following four competitors:

- **0-RA:** A Gaussian process with the true covariance function C_0, which provides the best possible fit, but scales as $\mathcal{O}(n^3)$.
- **1-RA F:** A "fast" 1-RA with fixed $r = 240$ and increasing $J = n/240$, which scales as $\mathcal{O}(n)$.
- **1-RA S:** A "slow" 1-RA with fixed $J = 64$ and increasing $r = n/64$, which scales as $\mathcal{O}(n^3)$.
- **M-RA:** An *M*-RA as described in Section 3.6, with $r = 30$, $J = 4$, and $M = \log_4(n/30)$, which scales as $\mathcal{O}(n\log^2 n)$.

For all competitors, the true covariance function (including all parameters) is assumed known, and we use the log-likelihood (at the true parameters) implied by each competitor as a measure of how well that competitor approximates this true covariance. The log-likelihood is equivalent to the log-score, which is a strictly proper scoring rule in the sense that it is maximized in expectation by the true model (e.g., Gneiting and Katzfuss 2014). This means that, on average, the 0-RA will have the highest possible log-score.

The results of these experiments with increasing sample size (averaged over the five datasets) are shown in Panels (a)–(c) of Figure 4. The computation times scale roughly as expected. We extrapolated the computation times of the 0-RA and 1-RA S for values of n for which the simulation machine ran out of memory, but were able to compute the exact log-likelihoods for the 0-RA using the Durbin–Levinson algorithm for up to $n \approx 500{,}000$. The *M*-RA and the 1-RA F had similar computation times, with the latter becoming slightly faster for very large n. The log-scores
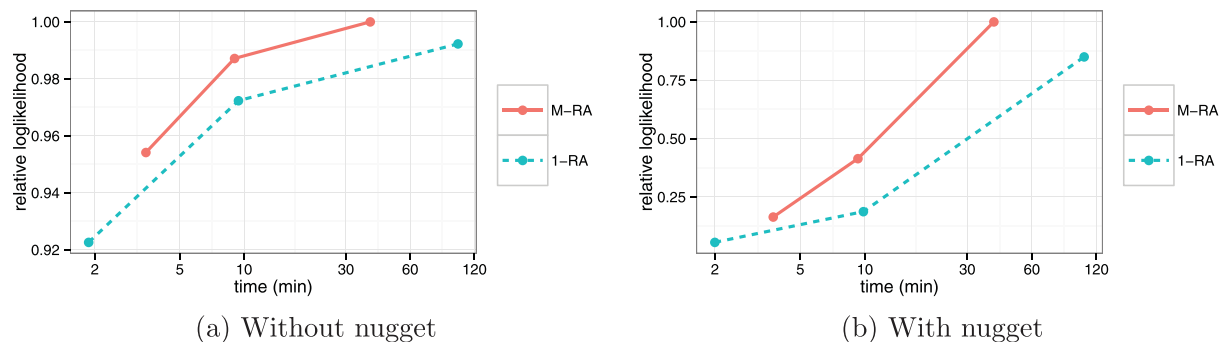

(a) Without nugget

(b) With nugget

Figure 5. For the simulation study in two dimensions, comparison of the *M*-RA (with varying M) to the 1-RA (with varying r). Note that the time axes are on a log scale. The log-likelihoods are all scaled relative to the log-likelihood of the 8-RA.

of the M-RA appear to be getting closer to those of the (computationally infeasible) 0-RA and 1-RA S with increasing n, while the log-scores of the 1-RA F become increasingly worse relative to the optimum.

For the largest sample size considered ($n = 1,966,080$), we further investigated computation times and log-scores for different versions of the M-RA (with $r = 30$ and $M = 2, 4, 8$) and the 1-RA (with r between 60 and 960). The results are shown in Panel (d) of Figure 4. The 2-RA and the 4-RA were roughly 8.7 and 11.8, respectively, faster than the fastest 1-RA with an equal or greater log-score. None of the 1-RAs achieved a log-score as high as the 8-RA.

We also conducted a simulation study in two-dimensional space. We considered $n = 3,211,264$ observations with an exponential covariance function with scale 0.3 and variance 1 on a regular grid on the unit square $\mathcal{D} = [0, 1]^2$. Using the function RFsimulate in the R package RandomFields (Schlather et al. 2015), we simulated five datasets without a nugget, and five datasets with a nugget consisting of Gaussian white noise with variance 0.05. For both settings, we compared the M-RA with $r = 49$ and $M = 2, 4, 8$ to the 1-RA with r between 49 and 784. The averaged results are shown in Figure 5. The relative performance of the two methods is similar to the one-dimensional case in Figure 4(d), but in the case without a nugget even fast approximations with small r or M achieve a relatively high log-likelihood.

4.2. Analysis of Total Precipitable Water

We also applied our methodology to $n = 271,014$ measurements of total precipitable water (TPW) made by the Microwave Integrated Retrieval System (MIRS) satellites between 2 a.m. and 3 a.m. UTC on February 1, 2011, over a region covering the United States. The measurements are shown in Panel (a) of Figure 6. The data are noisy and hourly datasets exhibit large gaps, which means that prediction of the true underlying TPW field is necessary at both observed and unobserved locations. Currently, an ad hoc operational version of such a gap-filled product is sent to National Weather Service offices, where it is used to track the movement of water vapor in the atmosphere and to detect conditions that can lead to heavy precipitation (see Kidder and Jones 2007; Forsythe et al. 2012, for more details).

We extended our methodology slightly to accommodate the fact that the TPW observations contain measurement error. We assumed that the observations were

$$z(\mathbf{s}_i) = y_M(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \quad i = 1, \ldots, n,$$

where $y_M(\cdot)$ is the M-RA as before, and for simplicity we assumed that we have spatially independent measurement error, $\epsilon(\mathbf{s}_i) \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$. In this case, parameter inference and prediction (of $y(\cdot)$, not $z(\cdot)$) can proceed as before, except that we needed to set $\mathbf{\Sigma}_{j_1,\ldots,j_M} = v_M(\mathcal{S}_{j_1,\ldots,j_M}, \mathcal{S}_{j_1,\ldots,j_M}) + \sigma_\epsilon^2 \mathbf{I}$ below (5).

We compared the proposed M-RA (with $M = 6$) to the 1-RA (full-scale approximation) and to a block-independent approximation (e.g., Stein 2014), which simply divides the domain into subregions and treats the process as independent between subregions. This can be viewed as a special case of the 1-RA with zero knots at resolution $m = 0$. The 6-RA had varying J_m at

different resolutions m, $(J_1, \ldots, J_6) = (2, 2, 4, 8, 8, 16)$, with an average number of 16.45 knots per region. The 1-RA had 1024 subregions with an average of 264.59 knots per region, and the block approximation had 256 subregions with an average of 1054.53 observations per region. After subtracting a constant mean, some exploratory analyses showed that a Matérn covariance with smoothness parameter 1.5 fit the data well, and so all methods used were approximating a covariance of the form,

$$C_0(\mathbf{s}_1, \mathbf{s}_2) = \sigma^2 \mathcal{M}_{1.5}(\|\mathbf{s}_1 - \mathbf{s}_2\|/\kappa),$$

where $\mathcal{M}_{1.5}$ is given in (14).

We first estimated the unknown parameters σ^2, κ, and σ_ϵ^2 by numerically maximizing the log-likelihood functions of the three approximation methods, and the resulting estimates and maximum log-likelihood values are given in Table 2. Then, using the estimated parameters, we computed the posterior distribution of the underlying TPW field on a regular $0.25° \times 0.25°$ latitude/longitude grid of size 24,805 over the domain. Marginal summaries (posterior means and posterior standard deviations) are shown in Figure 6.

Finally, we compared the posterior predictive distributions for three sets of 5000 randomly selected held-out test data (to evaluate short-range predictions) and for three randomly selected held-out test regions of size $5° \times 5°$ (to evaluate long-range predictions). As the true TPW values are unknown, we compared the predictions to the observations, considering the mean square prediction error (RMSPE) and the mean continuous rank probability score (CRPS). The CRPS is a strictly proper scoring rule that quantifies the fit of the entire predictive distribution (i.e., for a normal distribution, the mean, and the variance) to the data, and it is on the same scale as the observations (see, e.g., Gneiting and Katzfuss 2014).

As the block-independent long-range predictions were poor, we also carried out local kriging using the parameter estimates from the block-independent approximation. For every prediction location, the local-kriging predictions were based only on the data at the 20 nearest observed locations. The computation times for each test set were between 150 and 550 sec. Note also that the M-RA methods provide the joint posterior predictive distribution at all prediction locations, while local kriging only provides marginal posterior predictive distributions at each prediction location.

Summarizing the comparison in Table 2, the first three methods have similar computation times, maximum log-likelihood values, short-range predictions, and they all produce slight artifacts in the posterior-standard-deviation plots in the right column of Figure 6 in areas with nearly zero uncertainty. However, the 6-RA produces by far the best long-range predictions. In the prediction plots in the left column of Figure 6, strong "blocky" artifacts are visible for both the 1-RA and the block approximation. These differences are important in many satellite-data applications, where large regions of missing data in hourly or daily data are very common due to satellite tracks and nonretrieval (e.g., because of heavy cloud coverage).

5. Conclusions and Future Work

We have presented the multi-resolution approximation (M-RA), a novel technique for approximating Gaussian processes with

(a) TPW measurements

(b) Posterior means for the 6-RA   (c) Posterior standard deviations for the 6-RA

(d) Posterior means for the 1-RA   (e) Posterior standard deviations for the 1-RA

(f) Posterior means for the block approx.   (g) Posterior standard deviations for the block approx.
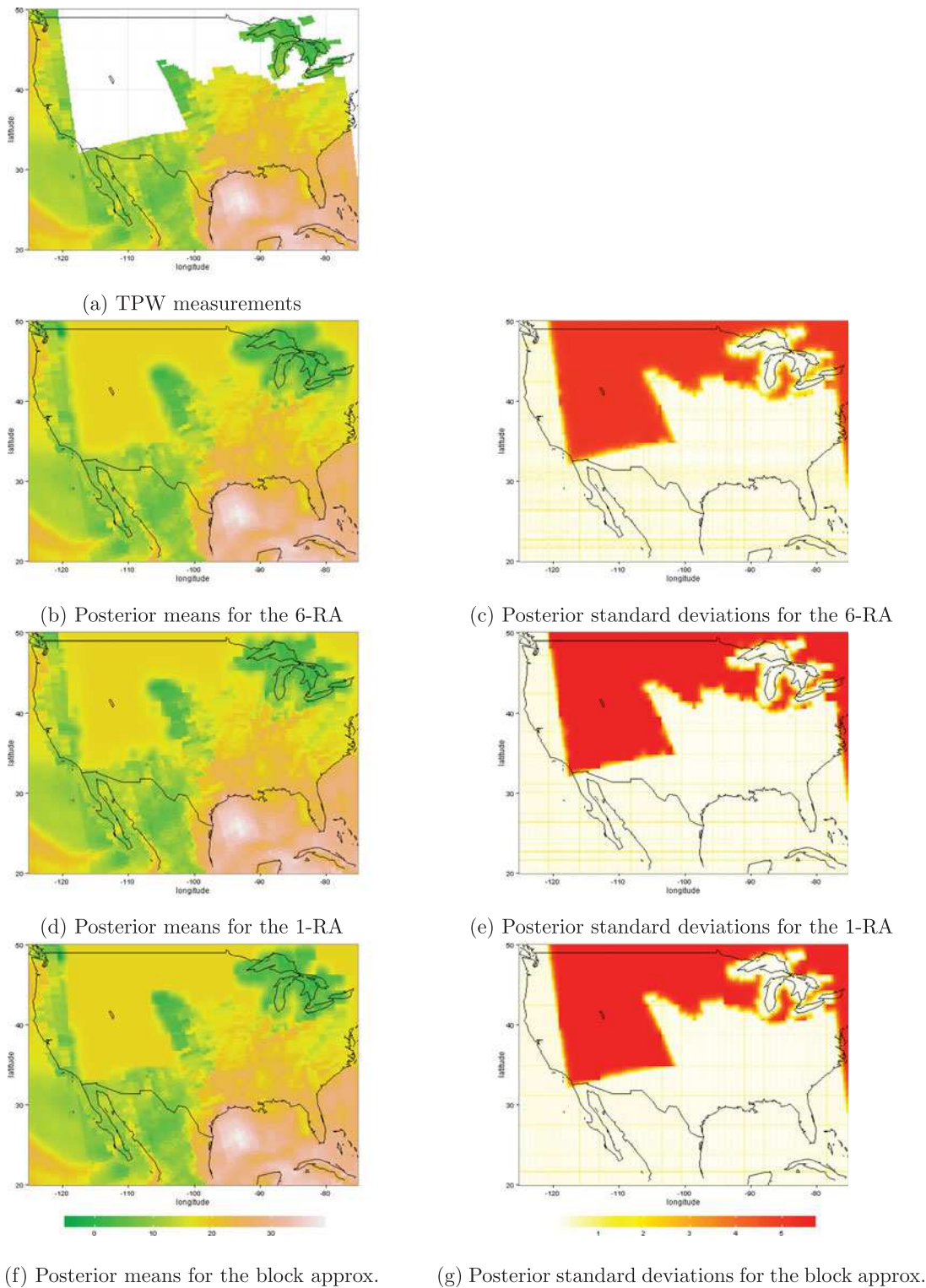
Figure 6. 271,014 measurements of total precipitable water (TPW), along with posterior predictive means and standard deviations of the true underlying TPW field on a 0.25° × 0.25° grid using three different methods. Color scales are in units of mm.

any covariance function. The M-RA is essentially a linear combination of many spatial basis functions at multiple resolutions. The precision matrix of the basis-function weights has a multiresolutional block-sparse structure, which allows scalable inference and distributed computations. Because the basis functions in our methodology are chosen optimally for a given covariance function, this can provide further insight on other multiresolution approaches in which basis functions are chosen in a more ad hoc way.

The M-RA compares favorably with the full-scale approximation of Sang, Jun, and Huang (2011), which is a current state-of-the-art method for large spatial data and can be viewed as a special case of the M-RA (with $M = 1$). Using theoretical results, a toy example, large simulated datasets, and a real-data application, we have shown that the M-RA can provide a better approximation at the same computational complexity and computation time as the 1-RA, or it can provide a similar approximation at a fraction of the computational time. It should also

Table 2. Results of the TPW analysis. block: block-independent approximation; local: local kriging based on 20 nearest neighbors; $\bar{r}$: average number of knots per region; time/lik.: average time (in seconds) per likelihood evaluation; loglik.: log-likelihood relative to the 6-RA log-likelihood; random: test set randomly sampled from observations; test regions: randomly selected test regions of size $5° \times 5°$; RMSPE: root mean square prediction error; CRPS: mean continuous rank probability score (lower is better).

		Maximum likelihood estimation					Random		Test regions	
	$\bar{r}$	time/lik.	$\hat{\sigma}^2$	$\hat{\kappa}$	$\hat{\sigma}_\epsilon^2$	loglik.	RMSPE	CRPS	RMSPE	CRPS
6-RA	16.45	98.32	26.31	0.74	0.29	1.00	0.56	0.22	3.24	1.72
1-RA	264.59	108.05	33.21	0.78	0.29	1.01	0.56	0.23	4.06	2.19
block	1054.53	94.61	31.22	0.73	0.29	1.00	0.56	0.23	4.51	2.38
local							0.56	0.22	3.79	2.00

be noted that our inference results for $M = 1$ provide an algorithm for parallel, distributed computations for inference in the full-scale approximation.

We are planning on providing user-friendly software that provides good default choices for the M-RA and that can be run on both desktop computers and on high-performance computing environments. Taking advantage of the distributed-memory architecture of the latter should in principle allow applying the M-RA to datasets with hundreds of millions of observations, as many satellite instruments are now able to produce on a daily basis.

The M-RA not only approximates the data covariance matrix, but it is a valid Gaussian process in its own right. Extensions to more complicated scenarios are therefore possible by embedding the M-RA process in a hierarchical model (e.g., Cressie and Wikle 2011). When the data measurement process is complex, the M-RA can be embedded in a hierarchical model that explicitly models the measurement process, and allows, for example, modeling non-Gaussian data, or fusing data from different measurement instruments.

Also of interest is a spatio-temporal version of the M-RA. Because it is possible to store and propagate the entire joint posterior predictive distribution, the M-RA could be extended to allow Kalman-filter-type inference in massive spatio-temporal state-space models (which is challenging for other sparse-precision approaches such as Lindgren, Rue, and Lindström 2011). In this sense, the M-RA might also provide an alternative to the ensemble Kalman filter (Evensen 1994; Katzfuss, Stroud, and Wikle 2016) in certain situations.

Appendix: Proofs

Proof of Proposition 1. For any set of locations $\mathcal{S} \subset \mathcal{D}$, $\mathbf{y}_M(\mathcal{S})$ in the form (3) is a linear combination of the vector consisting of all basis-function weights, which has a multivariate normal distribution, and so $y_M(\cdot)$ is a Gaussian process.

Further, note that $y_M(\cdot)$ in (1) is a sum of independent components, $\tau_m(\mathbf{s}) = E(\delta_m(\mathbf{s})|\delta_m(\mathcal{Q}^{(m)}))$, where $\delta_m(\cdot)$ is independent between regions $\mathcal{D}_{j_1,\dots,j_m}$. Starting with $\delta_0(\cdot) = y_0(\cdot)$, we can show iteratively for $m = 1, \dots, M-1$ using the law of total variance that, for any finite set $\mathcal{S}_{j_1,\dots,j_m} \subset \mathcal{D}_{j_1,\dots,j_m}$, the matrix

$$\text{var}\left(\delta_m\left(\mathcal{S}_{j_1,\dots,j_m}\right)\right) = \text{var}\left(\delta_{m-1}\left(\mathcal{S}_{j_1,\dots,j_m}\right)\right)$$
$$- \text{var}\left(E\left(\delta_{m-1}\left(\mathcal{S}_{j_1,\dots,j_m}\right)|\delta_{m-1}(\mathcal{Q}_{j_1,\dots,j_{m-1}})\right)\right)$$
$$= \text{var}\left(\delta_{m-1}\left(\mathcal{S}_{j_1,\dots,j_m}\right)|\delta_{m-1}(\mathcal{Q}_{j_1,\dots,j_{m-1}})\right)$$

is nonnegative definite. Thus, the covariance functions of the $\delta_m(\cdot)$, the $\tau_m(\cdot)$, and of $y_M(\cdot)$ are nonnegative definite. $\square$

Proof of Proposition 2. For $j_1, \dots, j_m = 1, \dots, J, m = 0, 1, \dots, M$, and $l = 0, \dots, m$, define

$$\boldsymbol{\mu}_{j_1,\dots,j_m}^l := E\left(\mathbf{y}_M\left(\mathcal{S}_{j_1,\dots,j_m}^P\right)|\mathbf{y}_M(\mathcal{S}), \mathcal{E}_{l-1}\right),$$

$$\boldsymbol{\Psi}_{i_1,\dots,i_m;j_1,\dots,j_m}^l := \text{cov}\left(\mathbf{y}_M(\mathcal{S}_{i_1,\dots,i_m}^P), \mathbf{y}_M\left(\mathcal{S}_{j_1,\dots,j_m}^P\right)|\mathbf{y}_M(\mathcal{S}), \mathcal{E}_{l-1}\right),$$

$$\mathbf{B}_{j_1,\dots,j_m}^{l,P} := \mathbf{b}_{j_1,\dots,j_l}\left(\mathcal{S}_{j_1,\dots,j_m}^P\right)$$

$$\mathbf{L}_{j_1,\dots,j_m}^l := \text{cov}\left(\mathbf{y}_M\left(\mathcal{S}_{j_1,\dots,j_m}^P\right), \mathbf{y}_M\left(\mathcal{S}_{j_1,\dots,j_l}\right)|\mathcal{E}_{l-1}\right)$$
$$= v_l\left(\mathcal{S}_{j_1,\dots,j_m}^P, \mathcal{S}_{j_1,\dots,j_l}\right),$$

$$\widetilde{\mathbf{B}}_{j_1,\dots,j_m}^{l,k} := \mathbf{B}_{j_1,\dots,j_m}^{k,P} - \mathbf{L}_{j_1,\dots,j_m}^l \boldsymbol{\Sigma}_{j_1,\dots,j_l}^{-1} \mathbf{B}_{j_1,\dots,j_l}^k,$$
$$k = 0, \dots, l-1.$$

Note that, for $l < M$, we have $\mathbf{L}_{j_1,\dots,j_m}^l = \mathbf{B}_{j_1,\dots,j_m}^{l,P} \mathbf{K}_{j_1,\dots,j_l} \mathbf{B}_{j_1,\dots,j_l}^l{}' + \widetilde{\mathbf{L}}_{j_1,\dots,j_m}^l$, where $\widetilde{\mathbf{L}}_{j_1,\dots,j_m}^l$ is a sparse block matrix with the only nonzero block being $\mathbf{L}_{j_1,\dots,j_m}^{l+1}$. Hence, it can be shown that $\widetilde{\mathbf{L}}_{j_1,\dots,j_m}^l \mathbf{V}_{j_1,\dots,j_l}^{-1} \mathbf{B}_{j_1,\dots,j_l}^k = \mathbf{L}_{j_1,\dots,j_m}^{l+1} \boldsymbol{\Sigma}_{j_1,\dots,j_{l+1}}^{-1} \mathbf{B}_{j_1,\dots,j_{l+1}}^k$. Using a variant of the Sherman–Morrison–Woodbury formula, we can also show that $\mathbf{K}_{j_1,\dots,j_l} \widetilde{\mathbf{A}}_{j_1,\dots,j_l}^{l,k} = \widetilde{\mathbf{K}}_{j_1,\dots,j_l} \mathbf{A}_{j_1,\dots,j_l}^{l,k}$. By applying (11), we therefore have

$$\widetilde{\mathbf{B}}_{j_1,\dots,j_m}^{l,k} = \mathbf{B}_{j_1,\dots,j_m}^{k,P} - \mathbf{B}_{j_1,\dots,j_m}^{l,P} \mathbf{K}_{j_1,\dots,j_l} \left(\mathbf{B}_{j_1,\dots,j_l}^l{}' \mathbf{V}_{j_1,\dots,j_l}^{-1} \mathbf{B}_{j_1,\dots,j_l}^k\right.$$

$$\left. - \mathbf{B}_{j_1,\dots,j_l}^l{}' \mathbf{V}_{j_1,\dots,j_l}^{-1} \mathbf{B}_{j_1,\dots,j_l}^l \widetilde{\mathbf{K}}_{j_1,\dots,j_l} \mathbf{B}_{j_1,\dots,j_l}^l{}' \mathbf{V}_{j_1,\dots,j_l}^{-1} \mathbf{B}_{j_1,\dots,j_l}^k\right)$$

$$- \widetilde{\mathbf{L}}_{j_1,\dots,j_m}^l \mathbf{V}_{j_1,\dots,j_l}^{-1} \mathbf{B}_{j_1,\dots,j_l}^k + \widetilde{\mathbf{L}}_{j_1,\dots,j_m}^l \mathbf{V}_{j_1,\dots,j_l}^{-1} \mathbf{B}_{j_1,\dots,j_l}^l \widetilde{\mathbf{K}}_{j_1,\dots,j_l} \mathbf{A}_{j_1,\dots,j_l}^{l,k}$$

$$= \mathbf{B}_{j_1,\dots,j_m}^{k,P} - \mathbf{B}_{j_1,\dots,j_m}^{l,P} \mathbf{K}_{j_1,\dots,j_l} \left(\mathbf{A}_{j_1,\dots,j_l}^{l,k} - \mathbf{A}_{j_1,\dots,j_l}^{l,l} \widetilde{\mathbf{K}}_{j_1,\dots,j_l} \mathbf{A}_{j_1,\dots,j_l}^{l,k}\right)$$

$$- \mathbf{L}_{j_1,\dots,j_m}^{l+1} \boldsymbol{\Sigma}_{j_1,\dots,j_{l+1}}^{-1} \mathbf{B}_{j_1,\dots,j_{l+1}}^k + \mathbf{L}_{j_1,\dots,j_m}^{l+1} \boldsymbol{\Sigma}_{j_1,\dots,j_{l+1}}^{-1} \mathbf{B}_{j_1,\dots,j_{l+1}}^l \widetilde{\mathbf{K}}_{j_1,\dots,j_l} \mathbf{A}_{j_1,\dots,j_l}^{l,k}$$

$$= \mathbf{B}_{j_1,\dots,j_m}^{k,P} - \mathbf{L}_{j_1,\dots,j_m}^{l+1} \boldsymbol{\Sigma}_{j_1,\dots,j_{l+1}}^{-1} \mathbf{B}_{j_1,\dots,j_{l+1}}^k$$

$$- \left(\mathbf{B}_{j_1,\dots,j_m}^{l,P} - \mathbf{L}_{j_1,\dots,j_m}^{l+1} \boldsymbol{\Sigma}_{j_1,\dots,j_{l+1}}^{-1} \mathbf{B}_{j_1,\dots,j_{l+1}}^l\right) \widetilde{\mathbf{K}}_{j_1,\dots,j_l} \mathbf{A}_{j_1,\dots,j_l}^{l,k}$$

$$= \widetilde{\mathbf{B}}_{j_1,\dots,j_m}^{l+1,k} - \widetilde{\mathbf{B}}_{j_1,\dots,j_m}^{l+1,l} \widetilde{\mathbf{K}}_{j_1,\dots,j_l} \mathbf{A}_{j_1,\dots,j_l}^{l,k},$$

which proves (13).

It is easy to see that the desired posterior predictive distribution is multivariate normal, $\mathbf{y}_M(\mathcal{S}^P)|\mathbf{y}_M(\mathcal{S}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Psi})$, and so spatial prediction amounts to finding the posterior mean and covariance matrix, $\boldsymbol{\mu}$ and $\boldsymbol{\Psi}$, respectively. To obtain these quantities, note that

we have from well-known properties of the multivariate normal distribution that

$$
\boldsymbol{\mu}_{j_1,\ldots,j_m}^{l} = E\left(\mathbf{y}_M\left(\mathcal{S}_{j_1,\ldots,j_m}^{P}\right) | \mathbf{y}_M\left(\mathcal{S}_{j_1,\ldots,j_l}\right), \mathcal{E}_{l-1}\right)
$$

$$
= \sum_{k=0}^{l-1} \mathbf{B}_{j_1,\ldots,j_m}^{k,P} \boldsymbol{\eta}_{j_1,\ldots,j_k} + \mathbf{L}_{j_1,\ldots,j_m}^{l} \boldsymbol{\Sigma}_{j_1,\ldots,j_l}^{-1}
$$

$$
\left(\mathbf{y}_M\left(\mathcal{S}_{j_1,\ldots,j_l}\right) - \sum_{k=0}^{l-1} \mathbf{B}_{j_1,\ldots,j_m}^{k} \boldsymbol{\eta}_{j_1,\ldots,j_k}\right)
$$

$$
= \mathbf{L}_{j_1,\ldots,j_m}^{l} \boldsymbol{\Sigma}_{j_1,\ldots,j_l}^{-1} \mathbf{y}_M(\mathcal{S}_{j_1,\ldots,j_l}) + \sum_{k=0}^{l-1} \widetilde{\mathbf{B}}_{j_1,\ldots,j_m}^{l,k} \boldsymbol{\eta}_{j_1,\ldots,j_k}.
$$

(A.1)

By the law of total expectation, we therefore have

$$
\boldsymbol{\mu}_{j_1,\ldots,j_m}^{l-1} = E\left(\boldsymbol{\mu}_{j_1,\ldots,j_m}^{l} | \mathbf{y}_M(\mathcal{S}), \mathcal{E}_{l-2}\right)
$$

$$
= \mathbf{L}_{j_1,\ldots,j_m}^{l} \boldsymbol{\Sigma}_{j_1,\ldots,j_l}^{-1} \mathbf{y}_M(\mathcal{S}_{j_1,\ldots,j_l}) + \sum_{k=0}^{l-2} \widetilde{\mathbf{B}}_{j_1,\ldots,j_m}^{l,k} \boldsymbol{\eta}_{j_1,\ldots,j_k}
$$

$$
+ \widetilde{\mathbf{B}}_{j_1,\ldots,j_m}^{l,l-1} \widetilde{\boldsymbol{\nu}}_{j_1,\ldots,j_{l-1}}
$$

$$
= \mathbf{L}_{j_1,\ldots,j_m}^{l} \boldsymbol{\Sigma}_{j_1,\ldots,j_l}^{-1} \mathbf{y}_M(\mathcal{S}_{j_1,\ldots,j_l}) + \widetilde{\mathbf{B}}_{j_1,\ldots,j_m}^{l,l-1} \widetilde{\mathbf{K}}_{j_1,\ldots,j_{l-1}} \widetilde{\boldsymbol{\omega}}_{j_1,\ldots,j_{l-1}}
$$

$$
+ \sum_{k=0}^{l-2} \widetilde{\mathbf{B}}_{j_1,\ldots,j_m}^{l,k} \boldsymbol{\eta}_{j_1,\ldots,j_k},
$$

(A.2)

and by the law of total covariance, we have

$$
\boldsymbol{\Psi}_{i_1,\ldots,i_m;j_1,\ldots,j_m}^{l-1} = E\left(\boldsymbol{\Psi}_{i_1,\ldots,i_m;j_1,\ldots,j_m}^{l} | \mathbf{y}_M(\mathcal{S}), \mathcal{E}_{l-2}\right)
$$

$$
+ \mathrm{cov}\left(\boldsymbol{\mu}_{i_1,\ldots,i_m}^{l}, \boldsymbol{\mu}_{j_1,\ldots,j_m}^{l} | \mathbf{y}_M(\mathcal{S}), \mathcal{E}_{l-2}\right)
$$

$$
= \boldsymbol{\Psi}_{i_1,\ldots,i_m;j_1,\ldots,j_m}^{l} + \mathrm{cov}
$$

$$
\left(\widetilde{\mathbf{B}}_{i_1,\ldots,i_m}^{l,l-1} \boldsymbol{\eta}_{i_1,\ldots,i_{l-1}}, \widetilde{\mathbf{B}}_{j_1,\ldots,j_m}^{l,l-1} \boldsymbol{\eta}_{j_1,\ldots,j_{l-1}} | \mathbf{y}_M(\mathcal{S}), \mathcal{E}_{l-2}\right)
$$

$$
= \boldsymbol{\Psi}_{i_1,\ldots,i_m;j_1,\ldots,j_m}^{l} + \widetilde{\mathbf{B}}_{j_1,\ldots,j_m}^{l,l-1} \widetilde{\mathbf{K}}_{j_1,\ldots,j_{l-1}} \widetilde{\mathbf{B}}_{j_1,\ldots,j_m}^{l,l-1}{}'
$$

$$
I((i_1,\ldots,i_m) = (j_1,\ldots,j_m)).
$$

(A.3)

The result (12) follows by starting with $\boldsymbol{\mu}_{j_1,\ldots,j_M}^{M}$ from (A.1) and $\boldsymbol{\Psi}_{j_1,\ldots,j_M}^{M} = \mathbf{V}_{j_1,\ldots,j_M}^{P} - \mathbf{L}_{j_1,\ldots,j_M}^{M} \boldsymbol{\Sigma}_{j_1,\ldots,j_M}^{-1} \mathbf{L}_{j_1,\ldots,j_M}^{M}{}'$, and iteratively applying (A.2) and (A.3) with $m = M$ for $l = M, \ldots, 0$. □

Supplementary Materials

All Julia code, R code to produce the plots, and the data for Section 4.2 are available online.

Acknowledgement

The author thanks Dorit Hammerling, Doug Nychka, Mikyoung Jun, Joe Guinness, Huiyan Sang, Suhasini Subba Rao, Valen Johnson, and two anonymous referees for helpful comments and suggestions. The author is also grateful to John Forsythe and Stan Kidder for providing the dataset in Section 4.2 and helpful advice.

References

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, 70, 825–848. [201,203,204]

Bevilacqua, M., Gaetan, C., Mateu, J., and Porcu, E. (2012), "Estimating Space and Space-Time Covariance Functions for Large Data Sets: A Weighted Composite Likelihood Approach," *Journal of the American Statistical Association*, 107, 268–280. [201]

Calder, C. A. (2007), "Dynamic Factor Process Convolution Models for Multivariate Space-Time Data With Application to Air Quality Assessment," *Environmental and Ecological Statistics*, 14, 229–247. [201]

Caragea, P. C., and Smith, R. L. (2007), "Asymptotic Properties of Computationally Efficient Alternative Estimators for a Class of Multivariate Normal Models," *Journal of Multivariate Analysis*, 98, 1417–1440. [201]

Chui, C. (1992), *An Introduction to Wavelets*, San Diego, CA: Academic Press. [201]

Cressie, N., and Johannesson, G. (2008), "Fixed Rank Kriging for Very Large Spatial Data Sets," *Journal of the Royal Statistical Society*, Series B, 70, 209–226. [201,202]

Cressie, N., and Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data*, Hoboken, NJ: Wiley. [201,212]

Curriero, F., and Lele, S. (1999), "A Composite Likelihood Approach to Semivariogram Estimation," *Journal of Agricultural, Biological, and Environmental Statistics*, 4, 9–28. [201]

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M., and Niemi, J. (2014), "Estimation and Prediction in Spatial Models With Block Composite Likelihoods Using Parallel Computing," *Journal of Computational and Graphical Statistics*, 23, 295–315. [201]

Evensen, G. (1994), "Sequential Data Assimilation With a Nonlinear Quasi-Geostrophic Model Using Monte Carlo Methods to Forecast Error Statistics," *Journal of Geophysical Research*, 99, 10143–10162. [212]

Forsythe, J. M., Dodson, J. B., Partain, P. T., Kidder, S. Q., and Haar, T. H. V. (2012), "How Total Precipitable Water Vapor Anomalies Relate to Cloud Vertical Structure," *Journal of Hydrometeorology*, 13, 709–721. [210]

Fuentes, M. (2007), "Approximate Likelihood for Large Irregularly Spaced Spatial Data," *Journal of the American Statistical Association*, 102, 321–331. [201]

Furrer, R., Genton, M. G., and Nychka, D. W. (2006), "Covariance Tapering for Interpolation of Large Spatial Datasets," *Journal of Computational and Graphical Statistics*, 15, 502–523. [201]

Gneiting, T., and Katzfuss, M. (2014), "Probabilistic Forecasting," *Annual Review of Statistics and Its Application*, 1, 125–151. [209,210]

Golub, G. H., and Van Loan, C. F. (2012), *Matrix Computations* (4th ed.), Baltimore, MD: JHU Press. [201]

Gramacy, R., and Lee, H. (2008), "Bayesian Treed Gaussian Process Models With an Application to Computer Modeling," *Journal of the American Statistical Association*, 103, 1119–1130. [205]

Halko, N., Martinsson, P., and Tropp, J. (2011), "Finding Structure With Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions," *SIAM Review*, 53, 217–288. [201]

Harville, D. A. (1997), *Matrix Algebra From a Statistician's Perspective*, New York: Springer. [207]

Henderson, H., and Searle, S. (1981), "On Deriving the Inverse of a Sum of Matrices," *SIAM Review*, 23, 53–60. [206]

Higdon, D. (1998), "A Process-Convolution Approach to Modelling Temperatures in the North Atlantic Ocean," *Environmental and Ecological Statistics*, 5, 173–190. [201,202]

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–417. [205]

Johannesson, G., Cressie, N., and Huang, H.-C. (2007), "Dynamic Multi-Resolution Spatial Models," *Environmental and Ecological Statistics*, 14, 5–25. [201]

Katzfuss, M. (2013), "Bayesian Nonstationary Spatial Modeling for Very Large Datasets," *Environmetrics*, 24, 189–200. [204,205]

Katzfuss, M., and Cressie, N. (2009), "Maximum Likelihood Estimation of Covariance Parameters in the Spatial-Random-Effects Model," in *Proceedings of the Joint Statistical Meetings*, Alexandria, VA: American Statistical Association, pp. 3378–3390. [201]

—— (2011), "Spatio-Temporal Smoothing and EM Estimation for Massive Remote-Sensing Data Sets," *Journal of Time Series Analysis*, 32, 430–446. [201]

—— (2012), "Bayesian Hierarchical Spatio-Temporal Smoothing for Very Large Datasets," *Environmetrics*, 23, 94–107. [201]

Katzfuss, M., and Hammerling, D. (2017), "Parallel Inference for Massive Distributed Spatial Data Using Low-Rank Models," *Statistics and Computing*, 27, 363–375. [202,206]

Katzfuss, M., Stroud, J. R., and Wikle, C. K. (2016), "Understanding the Ensemble Kalman Filter," *The American Statistician*, 70, 350–357. [212]

Kaufman, C. G., Schervish, M., and Nychka, D. W. (2008), "Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets," *Journal of the American Statistical Association*, 103, 1545–1555. [201]

Kidder, S. Q., and Jones, A. S. (2007), "A Blended Satellite Total Precipitable Water Product for Operational Forecasting," *Journal of Atmospheric and Oceanic Technology*, 24, 74–81. [210]

Lemos, R. T., and Sansó, B. (2009), "A Spatio-Temporal Model for Mean, Anomaly, and Trend Fields of North Atlantic Sea Surface Temperature," *Journal of the American Statistical Association*, 104, 5–18. [201]

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, 73, 423–498. [201,202,212]

Mardia, K., Goodall, C., Redfern, E., and Alonso, F. (1998), "The Kriged Kalman Filter," *Test*, 7, 217–282. [201]

McLeod, A. I., Yu, H., and Krougly, Z. (2007), "Algorithms for Linear Time Series Analysis: With R Package," *Journal of Statistical Software*, 23, 5. [208]

Nychka, D. W., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. R. (2016), "A Multi-Resolution Gaussian Process Model for the Analysis of Large Spatial Datasets," *Journal of Computational and Graphical Statistics*, 24, 579–599. [201,202]

Quiñonero Candela, J., and Rasmussen, C. (2005), "A Unifying View of Sparse Approximate Gaussian Process Regression," *Journal of Machine Learning Research*, 6, 1939–1959. [201,202]

Sang, H., and Huang, J. Z. (2012), "A Full Scale Approximation of Covariance Functions," *Journal of the Royal Statistical Society*, Series B, 74, 111–132. [202,204]

Sang, H., Jun, M., and Huang, J. Z. (2011), "Covariance Approximation for Large Multivariate Spatial Datasets With an Application to Multiple Climate Model Errors," *Annals of Applied Statistics*, 5, 2519–2548. [202,203,204,211]

Schlather, M., Malinowski, A., Menck, P. J., Oesting, M., and Strokorb, K. (2015), "Analysis, Simulation and Prediction of Multivariate Random Fields With Package Randomfields," *Journal of Statistical Software*, 63, 1–25. [210]

Shaby, B. A., and Ruppert, D. (2012), "Tapered Covariance: Bayesian Estimation and Asymptotics," *Journal of Computational and Graphical Statistics*, 21, 433–452. [201]

Sherman, J., and Morrison, W. (1950), "Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix," *Annals of Mathematical Statistics*, 21, 124–127. [202,206]

Snelson, E., and Ghahramani, Z. (2007), "Local and Global Sparse Gaussian Process Approximations," in *Artificial Intelligence and Statistics 11 (AISTATS)*. [202,204,205]

Stein, M. L. (2011), "2010 Rietz Lecture: When Does the Screening Effect Hold?," *Annals of Statistics*, 39, 2795–2819. [205]

—— (2014), "Limitations on Low Rank Approximations for Covariance Matrices of Spatial Data," *Spatial Statistics*, 8, 1–19. [204,210]

Stein, M. L., Chen, J., and Anitescu, M. (2013), "Stochastic Approximation of Score Functions for Gaussian Processes," *The Annals of Applied Statistics*, 7, 1162–1191. [201]

Stein, M. L., Chi, Z., and Welty, L. (2004), "Approximating Likelihoods for Large Spatial Data Sets," *Journal of the Royal Statistical Society*, Series B, 66, 275–296. [201]

Vecchia, A. (1988), "Estimation and Model Identification for Continuous Spatial Processes," *Journal of the Royal Statistical Society*, Series B, 50, 297–312. [201]

Woodbury, M. (1950), "Inverting Modified Matrices," Memorandum Report 42, Statistical Research Group, Princeton, NJ: Princeton University. [202,206]