Which parameterization of the Matérn covariance function?☆

Kesen Wang [a],[*], Sameh Abdulah [b], Ying Sun [a],[b], Marc G. Genton [a],[b]

[a] *Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia*
[b] *Extreme Computing Research Center (ECRC), King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia*

ARTICLE INFO

ABSTRACT

The Matérn family of covariance functions is currently the most popularly used model in spatial statistics, geostatistics, and machine learning to specify the correlation between two geographical locations based on spatial distance. Compared to existing covariance functions, the Matérn family has more flexibility in data fitting because it allows the control of the field smoothness through a dedicated parameter. Moreover, it generalizes other popular covariance functions. However, fitting the smoothness parameter is computationally challenging since it complicates the optimization process. As a result, some practitioners set the smoothness parameter at an arbitrary value to reduce the optimization convergence time. In the literature, studies have used various parameterizations of the Matérn covariance function, assuming they are equivalent. This work aims at studying the effectiveness of different parameterizations under various settings. We demonstrate the feasibility of inferring all parameters simultaneously and quantifying their uncertainties on large-scale data using the *ExaGeoStat* parallel software. We also highlight the importance of the smoothness parameter by analyzing the Fisher information of the statistical parameters. We show that the various parameterizations have different properties and differ from several perspectives. In particular, we study the three

most popular parameterizations in terms of parameter estimation accuracy, modeling accuracy and efficiency, prediction efficiency, uncertainty quantification, and asymptotic properties. We further demonstrate their differing performances under nugget effects and approximated covariance. Lastly, we give recommendations for parameterization selection based on our experimental results.

1. Introduction

The Matérn covariance function has been a longstanding tool for statisticians in spatial data analysis; see Porcu et al. (2023) for a comprehensive review of its application in spatial statistics, machine learning, physics, and numerous other relevant fields. Its purpose is to model the correlation between measurements at geospatial locations, allowing for predictions of unobserved values at new locations within the study region. Originally proposed by Matérn (1960) and later popularized by Handcock and Stein (1993), the Matérn covariance is stationary and isotropic, that is, distance-dependent and invariant under translations and rotations. It characterizes the spatial correlation between two random variables at locations $\mathbf{s}_i$ and $\mathbf{s}_j$, separated by a distance $h = \|\boldsymbol{h}\|$, where $\boldsymbol{h} = \mathbf{s}_i - \mathbf{s}_j$:

$$\mathrm{Cor}(\|\boldsymbol{h}\|; \boldsymbol{\theta}) = \frac{1}{2^{\nu-1}\Gamma(\nu)}(\kappa\|\boldsymbol{h}\|)^{\nu}\mathcal{K}_{\nu}(\kappa\|\boldsymbol{h}\|) \propto \int_{\mathbb{R}^d} e^{i\boldsymbol{h}^{\top}\mathbf{z}}(\|\mathbf{z}\|^2 + \kappa^2)^{-(2\nu+d)/2}\mathrm{d}\mathbf{z}, \quad \boldsymbol{\theta} = (\kappa, \nu)^{\top},$$

where d is the dimension of the spatial field, $\kappa > 0$ denotes the scale parameter, $\nu > 0$ represents the smoothness parameter, and $\mathcal{K}_{\nu}(\cdot)$ denotes the modified Bessel function of the second kind with order ν. The Matérn correlation function is independent of the field dimension d, which is integrated out from the spectral density $(\|\mathbf{z}\|^2 + \kappa^2)^{-(2\nu+d)/2}$. In addition, $\mathrm{Cor}(h; \boldsymbol{\theta})$ is standardized so that $\mathrm{Cor}(0; \boldsymbol{\theta}) = 1$. To make it a legitimate covariance function, statisticians often scale $\mathrm{Cor}(h; \boldsymbol{\theta})$ by a constant factor to account for non-unit variance. The Matérn covariance function is well-known for its flexibility which includes numerous covariance functions as special cases such as the exponential ($\nu = 1/2$), Whittle ($\nu = 1$), and squared-exponential ($\nu = \infty$) covariances. In the literature, variants of the Matérn covariance have been developed for different purposes and motivations. There are currently three ubiquitous parameterizations of the Matérn covariance function in the literature:

$$\mathcal{M}_1(h; \boldsymbol{\theta}_1) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}\left(\frac{h}{\beta}\right)^{\nu}\mathcal{K}_{\nu}\left(\frac{h}{\beta}\right) + \mathbb{1}_{h=0} \cdot \tau^2, \qquad \boldsymbol{\theta}_1 = (\sigma^2, \beta, \nu, \tau^2)^{\top}, \quad (1)$$

$$\mathcal{M}_2(h; \boldsymbol{\theta}_2) = \frac{\sqrt{\pi}\phi}{2^{\nu-1}\Gamma(\nu+\frac{1}{2})\alpha^{2\nu}}(\alpha h)^{\nu}\mathcal{K}_{\nu}(\alpha h) + \mathbb{1}_{h=0} \cdot \tau^2, \qquad \boldsymbol{\theta}_2 = (\phi, \alpha, \nu, \tau^2)^{\top}, \quad (2)$$

$$\mathcal{M}_3(h; \boldsymbol{\theta}_3) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}\left(\frac{2\sqrt{\nu}h}{\rho}\right)^{\nu}\mathcal{K}_{\nu}\left(\frac{2\sqrt{\nu}h}{\rho}\right) + \mathbb{1}_{h=0} \cdot \tau^2, \qquad \boldsymbol{\theta}_3 = (\sigma^2, \rho, \nu, \tau^2)^{\top}. \quad (3)$$

In variant $\mathcal{M}_1$, σ^2, β, and ν are the variance, range, and smoothness parameters, respectively. As for variant $\mathcal{M}_2$, α partially functions like an inverse range parameter and primarily affects low-frequency behaviors (Stein, 1999). Additionally, ϕ is the overall scale parameter that pertains to the high-frequency behavior of the spatial process. In variant $\mathcal{M}_3$, σ^2 plays the same role as in variant $\mathcal{M}_1$, and ρ resembles the functionality of β as the range parameter. The roles of the parameter ν, which controls the decay rate at high frequencies or, correspondingly, the field smoothness, is immutable in all three variants. Although these variants may seem quite different, their essence remains unchanged. There are scale or range parameters (which can be easily transformed from one to another as presented in Table 1) that measure the dependence range and smoothness parameters that govern the decay rate of the dependence. When $h = 0$, the three variants result in $\mathrm{Var}\{Z(\mathbf{s})\}$,

Table 1

Link functions for transformations in the $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ parameterizations of the Matérn covariance function.

↗	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
$\mathcal{M}_1$	$\sigma^2, \beta, \nu, \tau^2$	$\phi = \frac{\sigma^2\Gamma(\nu+\frac{1}{2})}{\pi^{1/2}\Gamma(\nu)\beta^{2\nu}}, \alpha = \frac{1}{\beta}, \tau^2$	$\sigma^2, \rho = 2\nu^{1/2}\beta, \nu, \tau^2$
$\mathcal{M}_2$	$\sigma^2 = \frac{\pi^{1/2}\phi\Gamma(\nu)}{\Gamma(\nu+\frac{1}{2})\alpha^{2\nu}}, \beta = \frac{1}{\alpha}, \nu, \tau^2$	$\phi, \alpha, \nu, \tau^2$	$\sigma^2 = \frac{\pi^{1/2}\phi\Gamma(\nu)}{\Gamma(\nu+\frac{1}{2})\alpha^{2\nu}}, \rho = \frac{2\nu^{1/2}}{\alpha}, \nu, \tau^2$
$\mathcal{M}_3$	$\sigma^2, \beta = \frac{\rho}{2\nu^{1/2}}, \nu, \tau^2$	$\phi = \frac{\sigma^2\Gamma(\nu+\frac{1}{2})(2\nu^{1/2})^{2\nu}}{\pi^{1/2}\Gamma(\nu)\rho^{2\nu}}, \alpha = \frac{2\nu^{1/2}}{\rho}, \nu, \tau^2$	$\sigma^2, \rho, \nu, \tau^2$

where $Z(\mathbf{s})$ represents a Gaussian random field (GRF) with locations $\mathbf{s}$, in differing forms. For instance, $\mathrm{Var}\{Z(\mathbf{s})\} = \sigma^2 + \tau^2$ in $\mathcal{M}_1$ and $\mathcal{M}_3$ but has a much more complicated expression in $\mathcal{M}_2$, as specified in the links to σ^2. Moreover, τ^2 represents the nugget effect due to the measurement error.

Regarding the associated applications, for instance, the correlation form of variant $\mathcal{M}_1$ was introduced and applied by Handcock and Stein (1993) to illustrate the idea of kriging in the Bayesian context. Abdulah et al. (2018b, 2023) have applied variant $\mathcal{M}_1$ to model large-scale geospatial data. It is also applied in the commonly used R package for spatial statistics *fields* (Nychka et al., 2021) and *geoR* (Ribeiro and Diggle, 2001) as the Matérn model. Hong et al. (2021) used variant $\mathcal{M}_1$ to assess prediction efficiency for large spatial datasets with approximated covariance functions. Variant $\mathcal{M}_2$ was employed in Stein (1999) to define the Matérn class, avoiding the problematic spectral density concentration at the origin happening in $\mathcal{M}_4(h; \boldsymbol{\theta}_2) = \phi(\alpha h)^\nu \mathcal{K}_\nu(\alpha h) + \mathbb{1}_{h=0} \cdot \tau^2$. In addition, variant $\mathcal{M}_2$ was also applied by Stein (1999) and Loh (2005) to study domain asymptotics for Matérn GRFs. Variant $\mathcal{M}_3$ was recommended by Handcock and Wallis (1994) and Stein (1999) for a more independent interpretation of ρ from ν compared to their counterparts in $\mathcal{M}_2$. Geoga et al. (2020) applied $\mathcal{M}_3$ for the log-likelihood optimization. Geoga et al. (2023) also used $\mathcal{M}_3$ to compute the Fisher information matrix. Furthermore, De Oliveira and Han (2022) studied the Fisher information pattern of ν with $\mathcal{M}_3$. In addition, $\mathcal{M}_3$ is employed in the R package *RandomFields* (Schlather et al., 2019), another popularly-used spatial statistics package, as one of the Matérn models.

As mentioned above, practitioners seem to use the three parameterizations rather arbitrarily and interchangeably with the assumption that they are identical. For instance, these variants can be easily transformed from one to another according to the link functions provided in Table 1. Thus, the three parameterizations can characterize the same field if the relationships provided in Table 1 are applied. In addition, the various parameterizations are redundant in the maximum likelihood estimation. Indeed, by the invariance property of maximum likelihood estimators (MLEs), we can transform the MLEs of one parameterization to those of another without losing optimality. However, the general interchangeability assumption of the three parameterizations is not quite right because numerous aspects require further consideration. First, the MLE is also a statistic that may have different statistical properties, such as asymptotic variance and correlation in various parameterizations for different domains and fields. Second, the differing parameter space of the three variants results in different shapes of the log-likelihood function (shown in Fig. 1 and in Figure S5 in the Supplementary Material) and, therefore, in differences in optimization performance. In particular, although the parameterizations $\mathcal{M}_1$ and $\mathcal{M}_3$ are mathematically similar, shapes of their log-likelihood functions are quite different near the optimal values (log-likelihood functions of $\mathcal{M}_3$ tend to be much flatter). In addition, the log-likelihood functions parameterized in $\mathcal{M}_2$ are quite different from those of $\mathcal{M}_1$ and $\mathcal{M}_3$. Therefore, a possible difference exists in the modeling and prediction efficiency for the three variants.

In this work, we examine the parameterizations $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ from the perspective of increasing-domain asymptotics and numerical optimization performance with the assessment tools implemented in *ExaGeoStat*, a high-performance parallel software for large-scale spatial statistics (Abdulah et al., 2018a). Among these benchmark functions, we rely on the Fisher information matrix to investigate the increasing-domain asymptotics of all parameter estimates, including the smoothness parameter, mean square prediction error (MSPE) for prediction performance, the mean loss of efficiency (MLOE) and the mean misspecification of the mean square error (MMOM) criteria

| $\mathcal{M}_1$, strong | $\mathcal{M}_2$, strong | $\mathcal{M}_3$, strong |
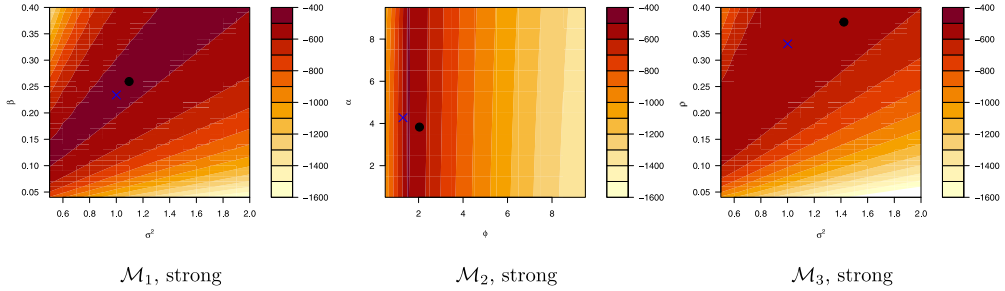
Fig. 1. Contour plots of the log-likelihood function of 1600 samples simulated from a strong exponential Gaussian field parameterized in $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ with $v = 0.5$. Contours of the log-likelihood functions for $v = 1$ tell a similar story and, thus, are omitted for concision. Contour plots for the 1600 samples simulated from weak and mediums exponential Gaussian fields are shown in Figure S5 in the Supplementary Material. The black dot in each figure represents the MLEs and the blue cross represents the true parameter values.

for modeling accuracy (Hong et al., 2021) and the total number of optimization iterations for modeling speed. Finally, we adopt the tile low-rank (TLR) covariance matrix approximation method, implemented by Abdulah et al. (2018b), to study the estimation accuracy of the three variants under an approximated covariance. Experiments under TLR settings are only conducted for $\mathcal{M}_1$ and $\mathcal{M}_3$ because $\mathcal{M}_2$ encounters numerical issues when compressing the covariance matrix in low-rank. All of the experiments are also repeated and jointly analyzed with nugget effects.

The paper is organized as follows. Section 2 presents a brief background of the *ExaGeoStat* software, and Section 3 introduces the embedded benchmarking functions of *ExaGeoStat*. Next, Section 4 analyzes the experimental results obtained from the benchmarking functions. Section 5 summarizes and studies the TLR covariance approximation method for the three variants. Lastly, Section 6 applies the three variants to a Saudi wind speed spatial dataset.

2. Large-scale spatial modeling and prediction

A significant challenge of handling large-scale spatial data is operating on the large covariance matrix of size $n \times n$, where n is the number of spatial locations. Computations involving matrices of large dimensions are usually of considerable complexity and become intractable as n becomes large. Specifically, evaluating the Gaussian log-likelihood function requires $O(n^2)$ memory space and $O(n^3)$ operations. Abdulah et al. (2018a) developed the geostatistical software *ExaGeoStat* to address this issue. It is a unified multicore high-performance computing software for large-scale geospatial data modeling and predictions on large-scale systems, including graphical processing units (GPUs) (Abdulah et al., 2018b). *ExaGeoStat* is a powerful tool for researchers working in geostatistics, providing advanced capabilities for large-scale spatial modeling and simulation that are not available in traditional software packages. *ExaGeoStat* adopts a three-layer software structure that includes a dynamic runtime system library, task-based parallel linear algebra solvers, and geospatial operations from bottom to top (Abdulah et al., 2023).

Upon its initial release, the developers implemented a stationary GRF data generator on a unit square based on $\mathcal{M}_1$. The data generator randomly selects a specific number of irregularly distributed locations within the unit square to calculate the distance matrix. Then it generates synthetic GRFs using the covariance matrix computed by applying $\mathcal{M}_1$ on the distance matrix. The synthetic data generator adopts parallel computing to split the large covariance matrix, which might be cumbersome, into smaller tiles and employs a column-wise filling procedure for each tile. The column-wise filling process operates in parallel on the tiles. In addition, *ExaGeoStat* performs all matrix operations in parallel by relying on state-of-the-art parallel linear algebra libraries to enhance computation efficiency significantly. The *ExaGeoStat* software also supports several approximation techniques in the covariance matrix, including independent block approximation

(IND), TLR (Abdulah et al., 2018b), and mixed-precision (Abdulah et al., 2021; Cao et al., 2022). The IND algorithm is based on the covariance approximation method discussed by Stein (2014), which attempts to decompose the large covariance matrix into blocks and annihilate off-diagonal blocks up to a prespecified bandwidth of the diagonal blocks. With TLR approximations, the authors aim to exploit the sparsity of the covariance matrix to reduce the memory footprint and decrease the execution time. With mixed-precision approximations, they aim to have more refined reduction techniques for the complexity of handling the covariance matrix by supporting lower-precision computation on the off-diagonal tiles.

3. Benchmarking functions

The objective of this study is to evaluate various variants of the Matérn covariance function using several metrics. In the subsequent subsections, we outline the specific metrics used in this study.

3.1. Maximum likelihood estimation

Using the Matérn form $\mathcal{M}_1$, $\boldsymbol{\theta}_1$ can be estimated by maximizing the log-likelihood function:

$$l(\boldsymbol{\theta}_1) = -\frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta}_1)| - \frac{1}{2}\mathbf{Z}^\top\boldsymbol{\Sigma}(\boldsymbol{\theta}_1)^{-1}\mathbf{Z} - \frac{n}{2}\log(2\pi), \tag{4}$$

where $\mathbf{Z}$ is a realization of a zero-mean GRF with covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}_1)$. As presented by Abdulah et al. (2018b), *ExaGeoStat*, the main software platform in this study, applies $\mathcal{M}_1$ on the distance matrix to generate $\boldsymbol{\Sigma}(\boldsymbol{\theta}_1)$ in a column-wise block parallel fashion. Analogously, in this work, we implement the generation function of $\boldsymbol{\Sigma}(\boldsymbol{\theta}_2)$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}_3)$ by applying the Matérn covariance functions $\mathcal{M}_2$ and $\mathcal{M}_3$ on the distance matrix. The synthetic data generator subsequently receives $\boldsymbol{\Sigma}(\boldsymbol{\theta}_2)$ or $\boldsymbol{\Sigma}(\boldsymbol{\theta}_3)$ to simulate realizations from the GRFs based on $\mathcal{M}_2$ or $\mathcal{M}_3$. Finally, the optimization function in *ExaGeoStat* maximizes $l(\boldsymbol{\theta}_2)$ and $l(\boldsymbol{\theta}_3)$ with the embedded model-based optimization algorithm **BOBYQA** presented by Powell (2009) to obtain MLEs of $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$, respectively. Using **BOBYQA** in *ExaGeoStat* avoids explicitly computing the first derivatives of the log-likelihood function, which can be cumbersome, especially when the sample size is large. To demonstrate this statement, we compute the first derivative of $l(\boldsymbol{\theta})$ based on the properties of matrix derivatives derived by Petersen et al. (2008) as follows:

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_k} = -\frac{1}{2}\,\text{tr}\left\{\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\frac{\partial\boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial\theta_k}\right\} + \frac{1}{2}\,\text{tr}\left\{\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{Z}\mathbf{Z}^\top\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\frac{\partial\boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial\theta_k}\right\}.$$

The expression above involves one call of matrix inversion and five calls of matrix multiplication. Matrix inversion and multiplication have high computational time complexity, which can take up to $O(n^2\log n)$ and $O(n^3)$ operations, respectively, rendering gradient-based algorithms non-scalable on large spatial datasets. Furthermore, the generation of $\frac{\partial\boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial\theta_k}$ relies on equations S1–S10 in the Supplementary Material, which can be slow when involving numerous calls of the modified Bessel function of the second kind especially in the computation of the derivative with respect to ν. Moreover, $\frac{\partial l(\boldsymbol{\theta})}{\partial\theta_k}$ is frequently unstable during the optimization, leading to problematic MLEs. Although gradient-based methods can take fewer iterations to converge given a reasonable starting point, they are overshadowed by the amount of time consumed for the gradient computation.

In addition, **BOBYQA** is an iterative local optimization approach subject to bounded constraints. The main rationale behind the mechanism of **BOBYQA** is to approximate a function $F(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$, through a quadratic form Q that satisfies $Q(\mathbf{y}_i) = F(\mathbf{y}_i)$, $i = 1, \ldots, m$, where $\mathbf{y}_i$ denotes an interpolation point. Then, each model update is realized by minimizing the Frobenius norm of the difference between the second-derivative matrices of Q, in symbols $\|\nabla^2 Q_{k+1} - \nabla^2 Q_k\|_\text{F}$. Updating from Q_k to Q_{k+1} only requires $O(n^2)$ operations if the number of interpolation points is $m = 2n + 1$ and calculating the change $Q_{k+1} - Q_k$ is limited to $O(m^2)$ operations. Therefore, **BOBYQA** is more computationally efficient than gradient-based algorithms consuming at least $O(n^3)$ operations, and more scalable on large datasets.

3.2. Fisher information matrix

The Fisher information matrix measures the amount of information on unknown parameters estimated from an observed sample and provides practical statistical insights into the uncertainty analysis of MLEs. Mardia and Marshall (1984) demonstrated that the Fisher information matrix could accurately approximate the covariance matrix of the MLEs even if the sample sizes are small. Abt and Welch (1998) illustrated that the covariance matrix of the limiting distribution of the MLEs for Gaussian processes is equal to the limit of the inverse Fisher information matrix, and it accurately approximates the variability and correlation between MLEs even in singularity conditions. To calculate the Fisher information matrix, we define a random variable X characterized by the m-dimensional parameter vector $\boldsymbol{\theta}$ with density function $f(x; \boldsymbol{\theta})$. The associated Fisher information matrix, $\mathbf{I}(\boldsymbol{\theta}) = \{I_{ij}(\boldsymbol{\theta})\}_{i,j=1,\dots,m}$, of X can be computed as follows:

$$I_{ij}(\boldsymbol{\theta}) = -\mathrm{E}\left[\left\{\frac{\partial}{\partial \theta_i}\log f(X; \boldsymbol{\theta})\right\}\left\{\log\frac{\partial}{\partial \theta_j}f(X; \boldsymbol{\theta})\right\}\right].$$

Geoga et al. (2020) recalled the asymptotic theory of MLEs, which states that if the smallest eigenvalue of $\mathbf{I}(\boldsymbol{\theta})$ tends to infinity as the sample size increases, then $\mathbf{I}(\hat{\boldsymbol{\theta}})^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to \mathcal{N}_m(\mathbf{0}, \mathbf{I}_m)$, where $\hat{\boldsymbol{\theta}}$ is the estimated parameter vector of the MLE. The asymptotic distribution of MLEs enables calculations of confidence intervals, which provide statistical tools to assess the estimation accuracy and quantify uncertainty. The Fisher information matrix of a GRF depends entirely on the covariance matrix and its first-order derivatives with respect to the parameters. Moreover, in a GRF, the mean μ is either assumed to be zero in common practice, as we typically model the residuals, or is empirically removed from the observations by subtracting the sample mean. Therefore, a sensible option is to exclude μ from the Fisher information matrix. Consequently, we can express the Fisher information matrix of a Matérn GRF, $\mathbf{I}(\boldsymbol{\theta}) = \{I_{ij}(\boldsymbol{\theta})\}_{i,j=1,2,3,4}$, in the form of taking the trace of a sequence of matrix multiplications (Mardia and Marshall, 1984):

$$\mathbf{I}_{ij}(\boldsymbol{\theta}) = \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{\Sigma}_j\right\}, \tag{5}$$

where $\boldsymbol{\Sigma}_i = \frac{\partial\boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial\theta_i}$ denotes the derivative of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ with respect to θ_i for $i = 1, 2, 3, 4$.

The computation of the Fisher information matrix, involving one call of matrix inversion and three calls of matrix multiplication, is usually costly when the sample size is large. To reduce complexity, Geoga et al. (2020) approximated the Fisher information matrix using hierarchical matrices, allowing for quasi-linear time complexity and memory space. The results from Geoga et al. (2020) revealed that the proposed hierarchical matrices accelerate the computing process while maintaining decent accuracy. However, this technique excessively relies on matrix approximations. Therefore, this work presents a parallel computing technique that computes the Fisher information precisely and controls the computational complexity.

To compute the Fisher information matrix of Matérn GRFs in parallel, we have to compute the first-order derivatives of $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ so that we can generate $\boldsymbol{\Sigma}_i$ in the same tile-based parallel fashion as we generate $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ as presented by Abdulah et al. (2018a). Section S1 in the Supplementary Material contains all the required first-order derivatives associated with $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$. In these equations, $\Psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ represents the digamma function. In addition, $\mathcal{K}'_\nu(x)$ denotes the derivative of the modified Bessel function of the second kind with respect to its argument x, and $\mathcal{K}_{\nu'}(x)$ denotes the first-order derivative taken with respect to its order ν. ExaGeoStat calculates $\mathcal{K}_\nu(x)$ using the GNU scientific library (Gough, 2009) and first-order derivatives of the modified Bessel function of the second kind with self-implemented functions based on the following identities:

$$\mathcal{K}'_\nu(x) = -\frac{1}{2}\{\mathcal{K}_{\nu-1}(x) + \mathcal{K}_{\nu+1}(x)\},$$
$$\mathcal{K}_{\nu'}(x) = \frac{\partial\mathcal{K}_\nu(x)}{\partial\nu} = \lim_{\Delta\to 0}\frac{\mathcal{K}_{\nu+\Delta}(x) - \mathcal{K}_\nu(x)}{\Delta}.$$

The first-order (and even higher-order) derivative of the modified Bessel function of the second kind with respect to ν does have closed-form expressions (González-Santander, 2018), and there

are series approximations that attempt to ease their numerical computation (Olver et al., 2010). However, we have found no existing well-compatible algorithms that perform this task. In Geoga et al. (2023), the authors provided an automatic differentiation method to compute $\frac{\partial \mathcal{K}_\nu(x)}{\partial \nu}$ by splitting the smoothness parameter ν into different categories and using the most efficient differentiation methodology for each. Nonetheless, the proposed software is currently only compatible with Julia. The lack of computational tools led us to implement a simple finite difference (FD) approach with an increment of $\Delta = 10^{-9}$, and we choose such a small tolerance because $\mathcal{K}_\nu(x)$, in most cases, is a smooth function of its order ν:

$$\frac{\partial \mathcal{K}_\nu(x)}{\partial \nu}\Big|_{\nu=n} = \frac{n!}{2} \sum_{i=0}^{n-1} \frac{(\frac{x}{2})^{i-n}}{i!(n-i)} \mathcal{K}_i(x), \quad n \in \mathbb{Z}^+, \qquad \text{Exact Integer Form (EXACT)}$$

$$\frac{\partial \mathcal{K}_\nu(x)}{\partial \nu} = \int_0^\infty t e^{-x\cosh(t)} \sinh(\nu t) \mathrm{d}t, \qquad \text{Integral Form (INTG)}$$

$$\frac{\partial \mathcal{K}_\nu(x)}{\partial \nu} \approx \mathcal{K}_\nu(x)\left\{ 1 - \frac{1}{2\nu} - \log\left(\frac{e \cdot x}{2\nu}\right) \right\}, \quad \nu \to \infty. \qquad \text{Asymptotic Form (ASYM)}$$

We aim to analyze GRFs on the unit square $[0, 1]^2$; thus, distances between observations are continuous on the interval $[0, 1]$. Then, we attempt to draw comparisons by discretizing the argument x as $(0.01, 0.3, 0.6, 0.9)$ and obtaining results across a range of values of the smoothness parameter ν based on the four approaches. Fig. 2 illustrates the comparison results. Fig. 2 precisely indicates that the finite difference approach is an accurate estimation of $\frac{\partial \mathcal{K}_\nu(x)}{\partial \nu}$ because the integral form and the exact integer form precisely overlap with the finite difference approach for all x. The asymptotic approximation converges to the finite difference approach reasonably rapidly. An additional observation is that asymptotic approximations converge slower when the argument x increases to the actual values. As presented in Fig. 2, the asymptotic approximation curves overlap with the finite difference and integral form curves at increasingly large values of ν.

Algorithm 1 Fisher Information Matrix Algorithm

Input: Set of locations ($\mathbf{s}$) and Matérn parameter vector ($\boldsymbol{\theta}_m$)
Output: Fisher information matrix $\mathbf{I}(\boldsymbol{\theta}_m)$
  Generate the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}_m)$ ▷ using $\mathcal{M}_1$, $\mathcal{M}_2$, or $\mathcal{M}_3$
  POTRF ($\boldsymbol{\Sigma}(\boldsymbol{\theta}_m)$)   ▷ Cholesky factorization $-\ \mathbf{L}\mathbf{L}^\top = \boldsymbol{\Sigma}(\boldsymbol{\theta}_m)$
  for $i = 1 : len(\boldsymbol{\theta}_m)$ **do**
    Generate the derivative covariance matrix $\boldsymbol{\Sigma}_i$ ▷ using $\mathcal{M}_1$, $\mathcal{M}_2$, or $\mathcal{M}_3$
    $\boldsymbol{\Sigma}_i \leftarrow$ TRSM ($\mathbf{L}^{-1}$, $\boldsymbol{\Sigma}_i$)   ▷ Triangular solve
    $\boldsymbol{\Sigma}_i \leftarrow$ TRSM ($\mathbf{L}^{-\top}$, $\boldsymbol{\Sigma}_i$)   ▷ Triangular solve
    for $j = i : len(\boldsymbol{\theta}_m)$ **do**
      Generate the derivative covariance matrix $\boldsymbol{\Sigma}_j$ ▷ using $\mathcal{M}_1$, $\mathcal{M}_2$, or $\mathcal{M}_3$
      $\boldsymbol{\Sigma}_j \leftarrow$ TRSM ($\mathbf{L}^{-1}$, $\boldsymbol{\Sigma}_j$)   ▷ Triangular solve
 $\boldsymbol{\Sigma}_j \leftarrow$ TRSM ($\mathbf{L}^{-\top}$, $\boldsymbol{\Sigma}_j$)   ▷ Triangular solve
      $\boldsymbol{\Sigma}_{temp} \leftarrow$ GEMM ($\boldsymbol{\Sigma}_i$, $\boldsymbol{\Sigma}_j$)   ▷ Matrix multiplication
      $\mathbf{I}_{ij}(\boldsymbol{\theta}_m)$ and $\mathbf{I}_{ji}(\boldsymbol{\theta}_m) \leftarrow \frac{1}{2}$ TRACE ($\boldsymbol{\Sigma}_{temp}$)

Algorithm 1 illustrates in detail the task-based parallel implementation of the Fisher information matrix. We rely on $\mathcal{M}_1$ to establish a basic algorithm architecture for the other parameterizations. The algorithm has two inputs, the set of locations ($\mathbf{s}$) and the Matérn parameter vector ($\boldsymbol{\theta}_m$). The algorithm begins with generating the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}_1)$ using $\mathcal{M}_1$ in parallel and performs a Cholesky factorization based on the tile-based matrices to obtain $\boldsymbol{\Sigma}(\boldsymbol{\theta}_1)^{-1}$. To estimate each Fisher matrix element, the parallel generation function fills in the entries for $\boldsymbol{\Sigma}_i$ or $\boldsymbol{\Sigma}_j$ with Equations (S1), (S2), and (S3) in the Supplementary Material. Next, the algorithm computes the full Fisher information matrix using a set of parallel linear algebra operations as shown by Eq. (5). We
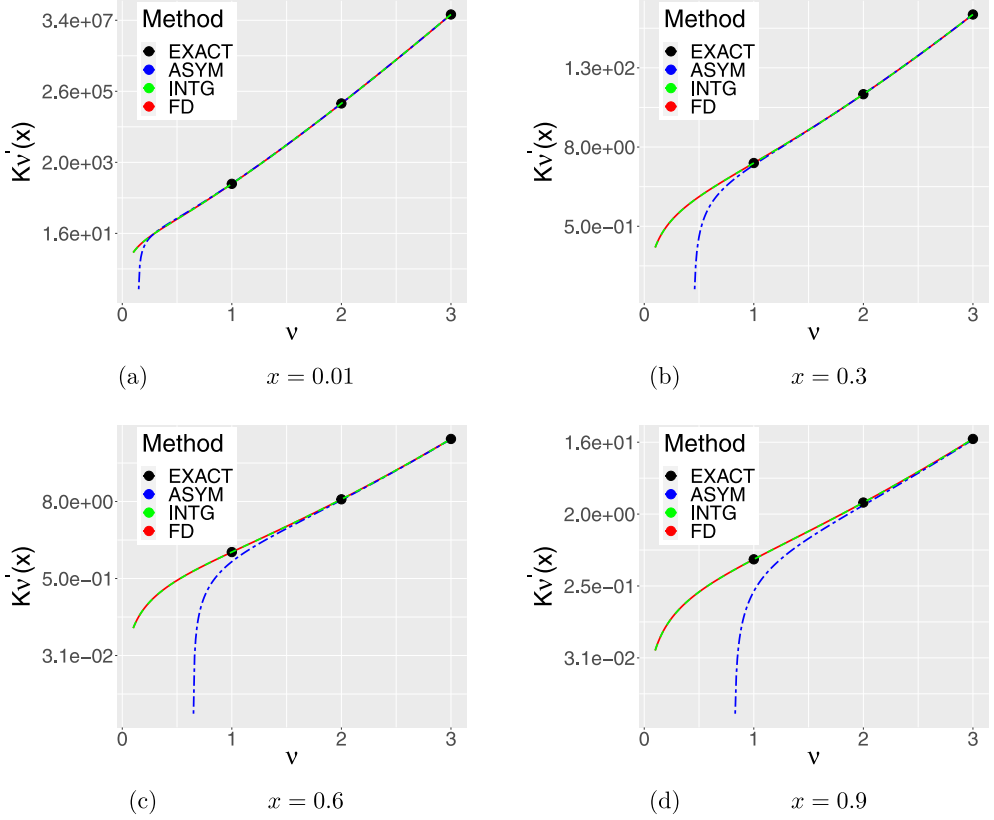
Fig. 2. Transformed $\log_2$ values of $\frac{\partial \mathcal{K}_\nu(x)}{\partial \nu}$ returned from the finite difference approach (FD, red), exact integer form (EXACT, black), asymptotic approximation (ASYM, blue), and integral form (INTG, green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

implemented two versions of Algorithm 1 for the other two parameterizations $\mathcal{M}_2$ and $\mathcal{M}_3$, together with their first-order derivative functions (S4)–(S9) to the existing implementation to compute their associated Fisher information matrices. The MLEs are obtained using the invariance property and the link functions provided in Table 1. Myung et al. (2005) illustrated that transformations of the Fisher information matrix are linear operations based on the Jacobian matrix; thus,

$$\mathbf{I}(\boldsymbol{\theta}_m) = \mathbf{J}_{k\to m}^\top \mathbf{I}(\boldsymbol{\theta}_k) \mathbf{J}_{k\to m}, \tag{6}$$

where $k, m = 1, 2, 3$ and $(\mathbf{J}_{k\to m})_{i,j} = \frac{\partial(\boldsymbol{\theta}_m)_i}{\partial(\boldsymbol{\theta}_k)_j}$ represents the Jacobian matrix for $i, j = 1, 2, 3, 4$. Then, we compute all the Jacobian matrices that enable transformations between $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ (see the Supplementary Material). The reverse transformations can be established by inverting the relationship presented in (6) and the computed Jacobian matrices.

The tile-based Fisher information matrix generation algorithm is adaptive to the parameter space. Thus, $\mathbf{I}(\boldsymbol{\theta})$ will be a 3×3 or 4×4 matrix, depending on the inclusion or exclusion of the nugget effect.

3.3. Mean squared prediction error

The given statistical model can predict missing measurements at new spatial locations through kriging using the posterior distribution of the missing observations based on the following:

$$\begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix} \sim \mathcal{N}_{m+n}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})), \text{ where } \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

The posterior mean, $\mathrm{E}(\mathbf{Z}_1|\mathbf{Z}_2 = \mathbf{z}_2) = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{z}_2$, can be used for the predictions of $\mathbf{Z}_1$, where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ represent the missing and observed values, respectively. The MSPE is:

$$\mathrm{MSPE} = \frac{1}{m} \sum_{i=1}^{m} \{\hat{Z}(\mathbf{s}_{0i}) - Z(\mathbf{s}_{0i})\}^2, \tag{7}$$

where $\hat{Z}(\mathbf{s}_{0i})$ is the predicted value at location $\mathbf{s}_{0i}$ and $Z(\mathbf{s}_{0i})$ is the observed value at location $\mathbf{s}_{0i}$. The MSPE directly accounts for the degree of deviation between the predicted and true values, a frequently used and standard cross-validation-based criterion to measure prediction errors and, therefore, quantify the prediction performance.

3.4. Data modeling accuracy

The concepts of loss of efficiency (LOE) and misspecification of the mean square error (MOM) were introduced by Stein (1999) to evaluate the impact of approximating the covariance on prediction error. LOE measures spatial prediction efficiency, while MOM quantifies the difference in MSPEs between a single prediction value calculated using the exact and approximated covariance values. Hong et al. (2021) extended these concepts to mean LOE (MLOE) and mean MOM (MMOM), which average LOE and MOM, respectively, across all locations with missing values. MLOE and MMOM provide a comprehensive assessment of the overall prediction accuracy and modeling error associated with the approximated covariance, and they are sensitive to inaccurate estimates and misspecified models. MLOE and MMOM are explicitly defined as $\mathrm{MLOE} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{LOE}(\mathbf{s}_{0i})$ and $\mathrm{MMOM} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{MOM}(\mathbf{s}_{0i})$, where $\mathrm{LOE}(\mathbf{s}_{0i}) = \mathrm{E}_t\{e_a^2(\mathbf{s}_{0i})\}/\mathrm{E}_t\{e_t^2(\mathbf{s}_{0i})\} - 1$ and $\mathrm{MOM}(\mathbf{s}_{0i}) = \mathrm{E}_a\{e_a^2(\mathbf{s}_{0i})\}/\mathrm{E}_a\{e_a^2(\mathbf{s}_{0i})\} - 1$. Here, $\mathbf{s}_{0i}$ represents the locations of the missing values, $e_a(\mathbf{s}_0)$ denotes the residuals computed using an approximated covariance, and $e_t(\mathbf{s}_0)$ represents the residuals computed with the true covariance. MLOE is non-negative since $\mathrm{E}_t\{e_t^2(\mathbf{s}_{0i})\} \leq \mathrm{E}_t\{e_a^2(\mathbf{s}_{0i})\}$ (Stein, 1999). MMOM is zero if parameter estimation is perfect; otherwise, it takes positive or negative values (Hong et al., 2021). Finally, $\mathrm{E}_a\{e_a^2(\mathbf{s}_0)\}$ is easy to compute, but evaluating $\mathrm{E}_t\{e_a^2(\mathbf{s}_0)\}$ is challenging.

Two methods were proposed by Stein (1999) for evaluating $\mathrm{E}_t\{e_a^2(\mathbf{s}_0)\}$. The first method involves estimating it by the conditional distribution of the estimate given the observations, which can be approximated as $\mathrm{E}_t\{e_a^2(\mathbf{s}_0)\} \approx \mathrm{E}_{\boldsymbol{\theta}}\left\{e_{\hat{\boldsymbol{\theta}}(\mathbf{Z})}^2(\mathbf{s}_0)|\mathbf{Z} = \mathbf{z}\right\} = \mathrm{E}_{\boldsymbol{\theta}}\{e_{\boldsymbol{\theta}}^2(\mathbf{s}_0)\} + \left\{e_{\hat{\boldsymbol{\theta}}(\mathbf{Z})}(\mathbf{s}_0) - e_{\boldsymbol{\theta}}(\mathbf{s}_0)\right\}^2$. Here, $\hat{\boldsymbol{\theta}}(\mathbf{Z})$ refers to the estimate, and $\boldsymbol{\theta}$ represents the true parameter value. The second method involves a resampling mechanism that repeatedly uses subsamples of $\mathbf{Z}$ to estimate $\boldsymbol{\theta}$, computes the prediction error for each subsample, averages these prediction errors, and adds them to $\mathrm{E}_t\{e_t^2(\mathbf{s}_0)\}$ to estimate $\mathrm{E}_t\{e_a^2(\mathbf{s}_0)\}$.

We utilized the covariance-based plug-in method proposed by Hong et al. (2021), which is based on Stein's approaches, to calculate both the true mean square prediction error (MSPE), $\mathrm{E}_t\{e_a^2(\mathbf{s}_0)\}$, and the approximated MSPE, $\mathrm{E}_a\{e_a^2(\mathbf{s}_0)\}$. To compute these, we employed the covariance matrix $\mathbf{K}_{\boldsymbol{\theta}} = \mathrm{Cov}_{\boldsymbol{\theta}}\{\mathbf{Z}, \mathbf{Z}^\top\}$, the covariance vector $\mathbf{k}_{\boldsymbol{\theta}} = \mathrm{Cov}_{\boldsymbol{\theta}}\{\mathbf{Z}, Z(\mathbf{s}_0)\}$, and the scalar variance $\mathrm{k}_{0,\boldsymbol{\theta}} = \mathrm{Var}_{\boldsymbol{\theta}}\{Z(\mathbf{s}_0)\}$. Specifically, we estimated $\mathrm{E}_t\{e_a^2(\mathbf{s}_0)\}$ as $\mathrm{E}_{\boldsymbol{\theta}}\left\{e_a^2(\mathbf{s}_0)\right\} \approx \mathrm{k}_{0,\boldsymbol{\theta}} - 2\mathbf{k}_{\boldsymbol{\theta}}^\top\mathbf{K}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{k}_{\hat{\boldsymbol{\theta}}} + \mathbf{k}_{\hat{\boldsymbol{\theta}}}^\top\mathbf{K}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{K}_{\boldsymbol{\theta}}\mathbf{K}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{k}_{\hat{\boldsymbol{\theta}}}$, and estimated $\mathrm{E}_a\{e_a^2(\mathbf{s}_0)\}$ as $\mathrm{k}_{0,\hat{\boldsymbol{\theta}}} - \mathbf{k}_{\hat{\boldsymbol{\theta}}}^\top\mathbf{K}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{k}_{\hat{\boldsymbol{\theta}}}$, where $\hat{\boldsymbol{\theta}}$ is the estimated parameter value. This matrix-based method allows for parallel computation, making it well-suited for large-scale datasets. We implemented this method in *ExaGeoStat*, which automatically considers the implemented covariance functions ($\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$) in the synthetic data generator and maximum likelihood estimators (MLEs) to construct true and approximated covariance matrices. The resulting MLOE and MMOM values are computed using the *mloe-mmom* command.

Table 2

Parameters specification in $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ for different correlation strengths for $\nu = 0.5$ and $\nu = 1$, and all with $\sigma^2 = 1$.

Dependence	$\nu = 0.5$				$\nu = 1$			
	β	ρ	ϕ	α	β	ρ	ϕ	α
Weak	0.0330	0.0467	9.6458	30.3030	0.0250	0.0500	800.0000	40.0000
Medium	0.1000	0.1414	3.1831	10.0000	0.0750	0.1500	88.8889	13.3333
Strong	0.2340	0.3309	1.3603	4.2735	0.1750	0.3500	16.3265	5.7143

4. Experimental results

In this section, we conduct a thorough evaluation of the Matérn covariance function's three parameterizations: $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$. Our objective is to assess the accuracy of parameter estimation for each variant, under exact and TLR approximate representations of the covariance matrix. Additionally, we evaluate the predictive performance of each covariance function by employing the MSPE, MLOE, and MMOM criteria. We also investigate the uncertainty quantification capabilities of each parameterization using the Fisher information matrix. Finally, we present our observations based on several experiments.

4.1. Experimental testbed

For the experimentation, we exploited the concept of weak, medium, and strong dependencies in GRF to discretize the field correlation strength to evaluate the three parametrizations of the Matérn covariance function. First, we specified the weak, medium, and strong dependence, using the effective range, which is the spatial lag for the correlation to drop to 5% (Stein, 1999). The effective range accounts for varying decay rates in the Matérn covariance function. The covariance function is expected to decrease at an increasingly slower rate as the correlation strengthens, addressing various scopes of dependence. Simulations in Abdulah et al. (2018b) revealed that the dependence range of GRFs affects the parameter estimation even in large-scale problems.

We applied the concept of effective range to determine the parameters in $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$. To specify the parameters, σ^2 is set to 1, and ν is set as 0.5 and 1, respectively, to address rough and smooth fields. Then, we computed β by numerically solving the equations, $\mathcal{M}_1(h; \boldsymbol{\theta}_1) = 0.05$ for $h = 0.1, 0.3, 0.7$. With all of the parameters specified in $\mathcal{M}_1$, we applied the link functions provided in Table 1 to specify parameters in $\mathcal{M}_2$ and $\mathcal{M}_3$. The nugget effect τ^2 is set to 0.1. Table 2 illustrates the parameters that describe identical Matérn GRFs parameterized in $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$. Table 2 reveals that the range and scale parameters (β, ρ) and α pertain to the correlation strength and, therefore, the decay rate of the covariance function. Moreover, β and ρ are direct measures of the dependence range. Increases in (β, ρ) can independently strengthen field correlation. In contrast, α, on the other hand, decreases as correlation strengthens. In addition, α cannot autonomously govern correlation strength. Decreasing values of ϕ are also imperative to increase the correlation strength.

4.2. Efficiency assessment

This section assesses the efficiency of the $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ covariance function variants concerning MLEs, the number of iterations to convergence, MLOE and MMOM criteria, and lastly MSPEs. We employed the synthetic data generator to simulate 300 replicates of GRFs of size 1600 in weak, medium, and strong fields with $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ to explore these measures. The synthetic data generator uses $[\{i - 0.5 \times \text{unif}(-0.4, 0.4)_i\}/\sqrt{n}, \{j - 0.5 \times \text{unif}(-0.4, 0.4)_i\}/\sqrt{n}]$ with $i, j = 1, \ldots, \sqrt{n}$ for the random locations, where unif denotes the uniform distribution and n is the sample size. Efficiency assessments are conducted on irregular grids because the behaviors of the metrics, such as MSPE, MLOE, MMOM, and MLEs, are similar on regular grids with dense observations. We used a 40-core Intel Cascade Lake machine with four V100 GPUs.

4.2.1. Parameter estimation accuracy

Fig. 3 presents the empirical description of the probabilistic distributions for the MLEs. We set the optimization tolerance to 10^{-5} and the tile size (*NB*) to be 50 for parallel execution. Tile size (NB) is tuned for the execution time performance. We adjusted the parameter search space by scaling the default optimization boundary in *ExaGeoStat*, (0.01, 5), by a constant factor to cover the MLEs adequately. For instance, we used the default boundaries if the true parameter was less than 2. For true parameter values between 2 and 20, we scaled the boundaries by 10. For the values within (20, 50) and (50, 500), we scaled the boundaries by 20 and 200, respectively. Lastly, for values greater than 500, we scaled the boundaries by 1000. The optimization boundaries are selected to properly contain the true parameter values and their respective MLEs. Because **BOBYQA** is a local optimization algorithm, each step will seek to find a more desirable position in the search space. As a result, **BOBYQA** will not end up on the boundaries if the MLEs are well-contained in the search space. Notice that in practice, we have no prior information on the true parameter values, but **BOBYQA** will always get stuck on the upper bounds (lower bounds can always be set to small values, like 0.01) if the MLEs are jointly or marginally larger than the upper bounds. Therefore, we recommend redoing the optimization with increased upper bounds under such circumstances until the MLEs are well-contained in the search space. In addition, we computed the difference ratios (DRs), |Mean − True Value|/True Value, and the approximated effective range (AER), $\mathcal{M}_i(h; \hat{\boldsymbol{\theta}}_i)$ for $h = 0.1, 0.3, 0.7$ and $i = 1, 2, 3$. In the Supplementary Material, parameter estimates on the Whittle Gaussian fields are plotted in Figure S1. The DRs calibrate differing parameter scales for fair comparisons and refer to the performance of point estimations. The AER measures the overall accuracy of the mean estimates.

For the non-nugget effect models, Fig. 3 (No-Nug) and Table S1 (first and third horizontal blocks separated by double horizontal lines in the Supplementary Material) demonstrate that estimates of $\mathcal{M}_3$ are the least accurate compared with their counterparts in $\mathcal{M}_1$ and $\mathcal{M}_2$. In particular, $\hat{\sigma}^2$ in $\mathcal{M}_3$ appears more biased than $\hat{\sigma}^2$ in $\mathcal{M}_1$. Also $\hat{\rho}$ is more variable and biased than $\hat{\beta}$. Although parameters in $\mathcal{M}_2$ are on different scales, Table S1 illustrates that $(\hat{\sigma}^2, \hat{\rho})$ are more biased than $(\hat{\phi}, \hat{\alpha})$. The biases of $(\hat{\sigma}^2, \hat{\rho})$ are due to flatter shapes of the associated log-likelihood functions near the maxima (leading the optimization algorithm to stop at non-optimal values) as shown in Figs. 1 and S5 in the Supplementary Material, which a more stringent convergence condition can resolve (i.e., 10^{-9}) at the expense of modeling speed. Table S3 (vertical block on the left separated by double vertical lines) also reveals that $\mathcal{M}_3$ provides undesirable AERs compared with $\mathcal{M}_1$ and $\mathcal{M}_2$, reflecting its less ideal overall estimation efficiency. $\mathcal{M}_2$ provides more refined AERs than $\mathcal{M}_1$ but performs poorly on point estimations of $(\hat{\phi}, \hat{v})$ in many cases. Furthermore, $(\hat{\phi}, \hat{\alpha})$ exhibits considerably higher volatility than $(\hat{\sigma}^2, \hat{\beta}, \hat{\rho})$, which can be observed through their large interquartile ranges (IQRs) shown in Figs. 3(d)–3(f) and S1d–S1f.

For the nugget effect models, the nugget effect significantly influences the estimates, but the effect varies in terms of $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$. From the experiments, the influence of the nugget effect is more crucial on $\mathcal{M}_1$ and $\mathcal{M}_2$. Fig. 3 (Nug) and Tables S1 (second and fourth horizontal blocks separated by double horizontal lines in the Supplementary Material), S3 (vertical block on the right separated by double vertical lines in the Supplementary Material) demonstrate that the nugget effect models of $\mathcal{M}_1$ and $\mathcal{M}_2$ both suffer from biased point estimates for $(\hat{\sigma}^2, \hat{\beta}, \hat{\phi}, \hat{\alpha})$, and undesirable AERs. In addition, $(\hat{\sigma}^2, \hat{\beta}, \hat{\phi})$ also becomes more variable in the presence of the nugget effect for $\mathcal{M}_1$ and $\mathcal{M}_2$. In contrast, parameter estimates in $\mathcal{M}_3$ maintain more robustness. In the nugget effect models, $\hat{\sigma}^2$ is less biased and less variable. $\hat{\rho}$ is less biased and about equally variable under the nugget effect. Impact of the nugget effect on $\hat{v}$ is similar across all parameterizations; $\hat{v}$ in all parameterizations tends to be more variable and biased under the nugget effect.

Fig. 4 confirms the conclusion of the nugget effect proposed by Stein (1999), stating that τ^2 should be easier to estimate in a smooth process. Fig. 4 also indicates that the IQRs of τ^2 decrease from an exponential Gaussian field to a Whittle Gaussian field. In addition, within the same field, the IQRs decline as the field correlation strengthens. The decrease in IQRs implies that τ^2 has a lower uncertainty and, therefore, more information in a smooth process, agreeing with the decreasing L^τ/τ^2 along with increasing v calculated on p. 186–187 in work by Stein (1999). Moreover, the estimations of τ^2 are less accurate in $\mathcal{M}_2$ compared with $\mathcal{M}_1$ and $\mathcal{M}_3$, which tends to have more
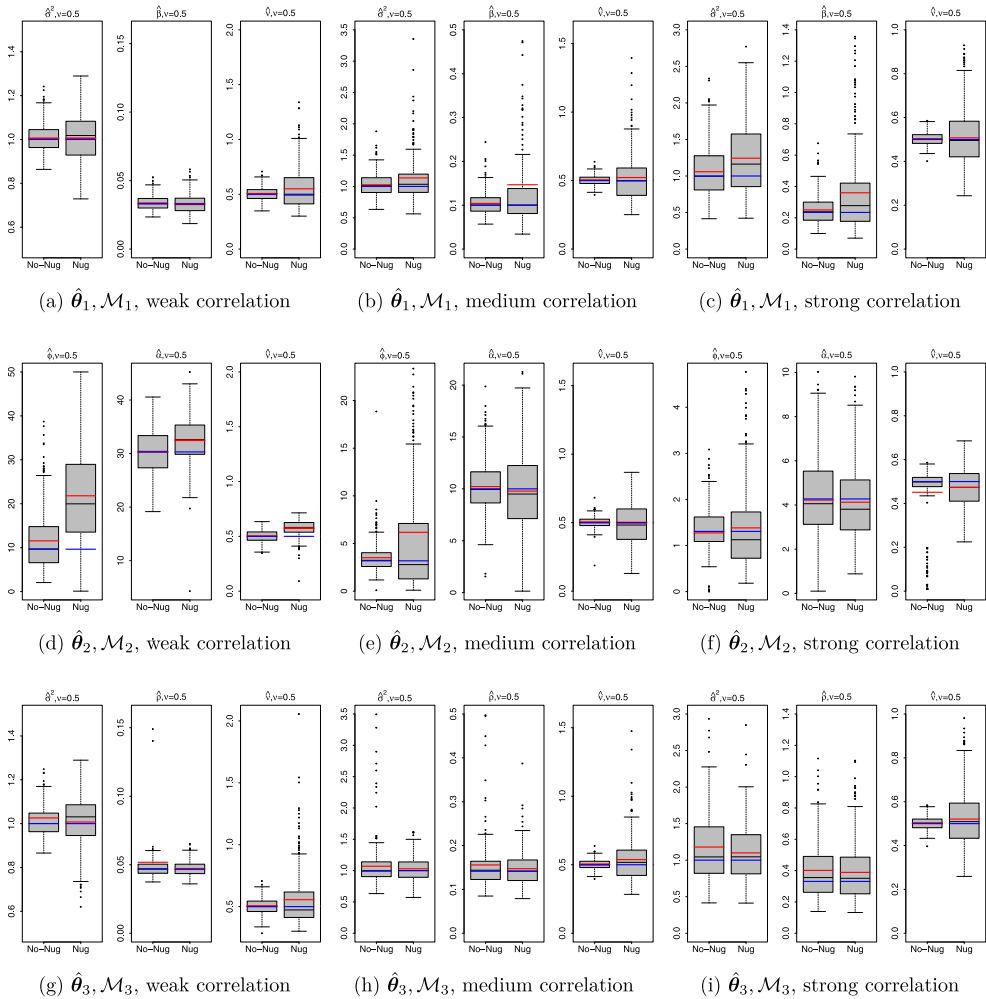
Fig. 3. Boxplots of MLEs on the exponential Gaussian fields ($\nu = 0.5$) using 300 replicates. Red lines represent the sample means; blue lines denote the true values. No-Nug denotes non-nugget effect models and Nug denotes nugget effect models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

significant deviations between the sample means and the true value. Lastly, IQRs decrease for $(\hat{\phi}, \hat{\alpha}, \hat{\nu})$ but increase for $(\hat{\sigma}^2, \hat{\beta}, \hat{\rho})$ as the field strengthens, indicating opposite behaviors of their uncertainties. Therefore, the estimation is more challenging for $(\hat{\sigma}^2, \hat{\beta}, \hat{\rho})$ but easier for $(\hat{\phi}, \hat{\alpha}, \hat{\nu})$ when the observations are correlated. Such differing behaviors of the estimates are due to variations in the data efficiency of their estimations, which we detail in Section 4.3.1 through their Fisher information.

The MLOE, which is small (i.e., $<10^{-3}$), does not differ among the three parameterizations for all the non-nugget effect models, indicating qualified prediction efficiency for all three parameterizations. Nugget effects increase the MLOE for all three parameterizations, incurring prediction efficiency loss, and they are comparable around 5×10^{-2}. We also calculated the MMOM criteria to scrutinize the modeling accuracy of $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$. Fig. 5 presents boxplots of MMOM obtained from the 300 replicates. For each replicate, we randomly single out 320 locations (20%) as missing
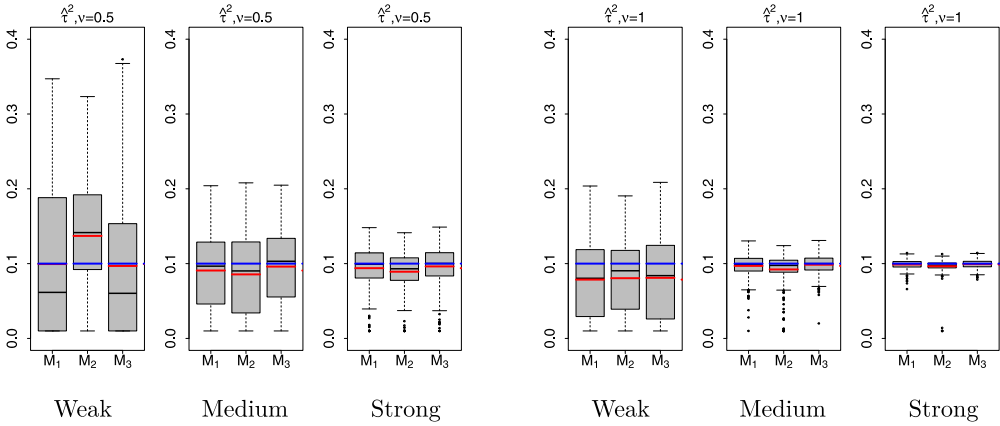
Fig. 4. Boxplots of MLEs for the nugget effect τ^2 on the exponential Gaussian fields ($v = 0.5$) (left three subfigures) and Whittle Gaussian fields ($v = 1$) (right three subfigures) using 300 replicates. Red lines represent the sample means; blue lines denote the true parameter values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

values to calculate $E_t\{e_a^2(\mathbf{s}_{0i})\}$ and $E_a\{e_a^2(\mathbf{s}_{0i})\}, i = 1, \ldots, 320$. The MMOM on the Whittle Gaussian fields displays similar patterns. Hence, we only provide the MMOM on the exponential Gaussian fields.

Fig. 5 indicates that the MMOM of $\mathcal{M}_2$ has numerous outliers and the largest IQRs across all field strengths, signifying frequent appearances of biased estimates. This result agrees with the high volatility issue of $\hat{\boldsymbol{\theta}}_2$ presented in Figs. 3(d)–3(f). Large variability of the MMOM can potentially diminish the reliability of the MSPE as a measure of prediction performance because the approximated MSPE is likely to differ from the true MSPE. Therefore, cross-validation approaches based on mean squared errors may fail to capture the actual performance of $\hat{\boldsymbol{\theta}}_2$ in practice. Nugget effects influence MMOMs, and the effects are more critical on $\mathcal{M}_2$. First, nugget effects generally increase the IQRs of MMOMs, increasing the possibility for the approximated MSPE to deviate from the true MSPE. Wider IQRs of MMOMs can be ascribed to the higher volatility of $\hat{\boldsymbol{\theta}}$ in the presence of nugget effects, as presented in Fig. 3. Second, the MMOMs of $\mathcal{M}_2$ with nugget effects are more inclined to significant IQR increases and are more likely to be nonzero, implying that the approximated MSPE can frequently overestimate or underestimate the true prediction error. In contrast, MMOMs of $\mathcal{M}_1$ and $\mathcal{M}_3$ are zero-centered with milder IQRs' increases for all field strengths, indicating more stable estimates and a reliable approximated MSPE. Consequently, modeling $\mathcal{M}_2$ with nugget effects is more likely to result in misspecified models, especially for weak and strong fields.

4.2.2. Prediction

In this part of the experiment, we calculate the MSPEs to quantify the prediction performance of $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ under various scenarios. We visualize empirical distributions of the MSPEs of the 300 replicates in Figs. 6 and S2 in the Supplementary Material.

Figs. 6 and S2 demonstrate that the prediction performance is stable and comparable for $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ for all settings. The MSPEs have approximately identical distributions with relatively narrow IQRs and few outliers. Complicated parameterizations, such as $\mathcal{M}_2$ and $\mathcal{M}_3$, do not produce improved prediction errors. However, as mentioned above, the MSPE of $\mathcal{M}_2$ is not as representative of the prediction performance as the MSPEs of $\mathcal{M}_1$ and $\mathcal{M}_3$. Although the three parameterizations have varying behaviors of the MMOM, identical MSPE values are possible because the MSPE is only a superficial measure. Hong et al. (2021) demonstrated similar results. In addition, within the same field, there is a downward trend of the MSPE regarding the field strength. In different fields, the
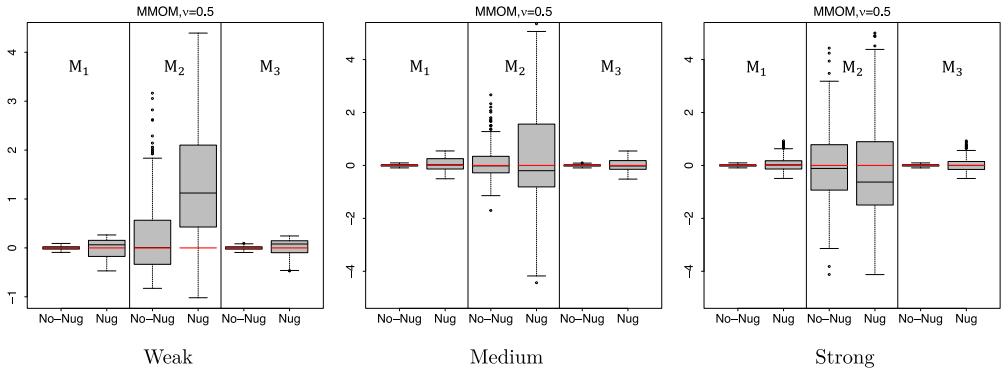
Fig. 5. Boxplots of MMOM criteria using 300 replicates on the exponential Gaussian fields ($\nu = 0.5$). For all subfigures above, $\mathcal{M}_1$ is in the first block, and $\mathcal{M}_2$ and $\mathcal{M}_3$ fill in the second and third blocks, respectively. Vertical lines separate the blocks. Red lines represent the expected MMOM value when true parameter values are used. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
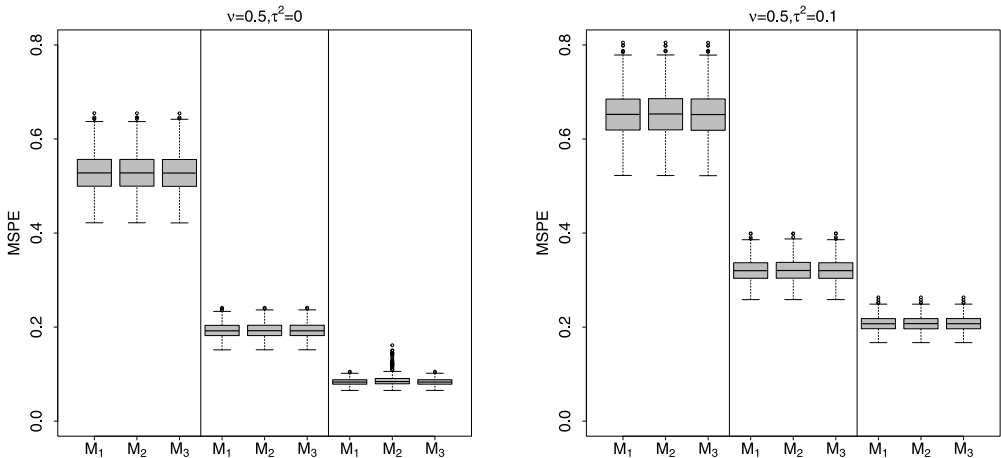
Fig. 6. Boxplots of the MSPEs using 300 replicates for non-nugget effect models ($\tau^2 = 0$) and nugget effect models ($\tau^2 = 0.1$) under weak (First block), medium (second block), and strong (third block) correlations, separated by vertical lines in each sub-figure, on the exponential Gaussian fields ($\nu = 0.5$).

MSPE decreases as the spatial process becomes smoother (exponential to Whittle Gaussian field). These results are consistent with the property of the MSPE of best linear unbiased predictor (BLUP) reported in Zimmerman and Cressie (1992), which stated that the BLUP is the most efficient (lowest MSPE) when the spatial correlation is strong. Nugget effects increase prediction errors due to the addition of noise in the spatial process.

The estimation accuracy of (σ^2, ϕ) does not affect the MSPE. In Section 3.3, spatial predictions are calculated as $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{z}_2$, which can be rewritten as $\tilde{\boldsymbol{\Sigma}}_{12}\tilde{\boldsymbol{\Sigma}}_{22}^{-1}\mathbf{z}_2$, where $\boldsymbol{\Sigma} = \sigma^2\tilde{\boldsymbol{\Sigma}}$ in $\mathcal{M}_1$ and $\mathcal{M}_3$ or $\boldsymbol{\Sigma} = \phi\tilde{\boldsymbol{\Sigma}}$ in $\mathcal{M}_2$. Nonetheless, variations of their values do crucially influence the MSPE. For instance, if the rest of the parameters are fixed, increases in σ^2 or ϕ yield a larger MSPE. In contrast, neither variations of α nor its estimation accuracy can have significant effects on the MSPE, which has been accounted for by Stein (1999), stating that high-frequency parameters primarily affect spatial interpolations. As for the range parameters, the MSPE is more sensitive to β than ρ. It takes a sharper increase in ρ for the field parameterized in $\mathcal{M}_3$ to attain the same strength as in $\mathcal{M}_1$. In
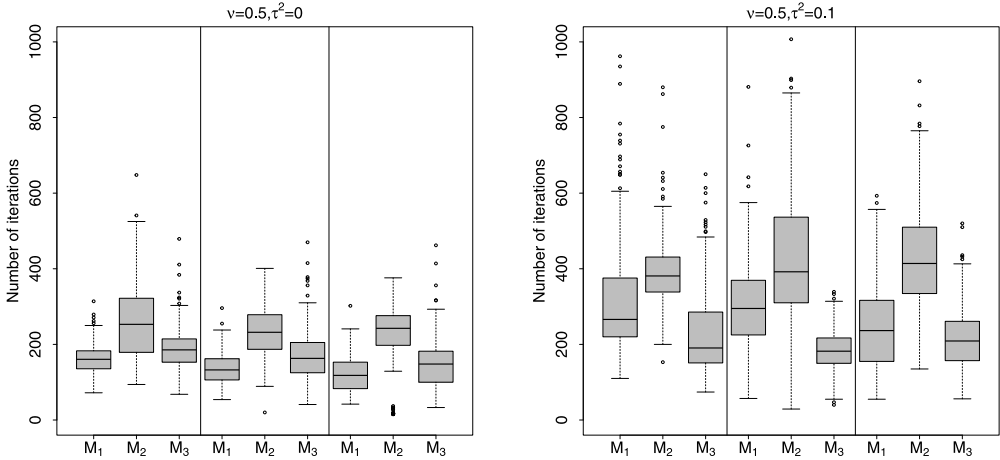
Fig. 7. Boxplots of the convergent iterations obtained using the 300 replicates for non-nugget effect models ($\tau^2 = 0$) and nugget effect models ($\tau^2 = 0.1$) under weak (first block), medium (second block), and strong (third block) correlations, separated by vertical lines in each sub-figure, on the exponential Gaussian fields ($\nu = 0.5$).

other words, the same amount of MSPE variation requires a more dramatic change in ρ than β. A higher sensitivity of β should always hold except for extremely rough fields where $\nu < 0.25$. The 0.25 bound comes from the link function between β and ρ provided in Table 1, $\rho = 2\nu^{1/2}\beta$.

4.2.3. Evaluating convergence rate

We visualized the number of iterations the optimization algorithm requires to converge and found the optimal values of the underlying parameters with a given tolerance. We henceforth refer to the number of iterations to converge as the convergent iterations for concision. Moreover, we only display the convergent iterations on the exponential Gaussian field because they exhibit the same patterns on the Whittle Gaussian field.

Figs. 7 and S3, in the Supplementary Material, indicate that $\mathcal{M}_2$ has the most significant convergent iterations. This low convergence rate can be anticipated from its ample parameter search space. Nevertheless, the large parameter space is inevitable for decent convergence of $(\hat{\phi}, \hat{\alpha})$. According to the theory of asymptotic normality, we have for some settings that $\hat{\phi} \sim \mathcal{N}(800, 596200)$ and $\hat{\alpha} \sim \mathcal{N}(40, 27.5)$ as calculated in Table S6. It is possible to accelerate the modeling process by reducing the search space at the risk of losing optimal values because the variances of $\hat{\boldsymbol{\theta}}_2$ are quite extraordinary. The other crucial factor that contributes to the slow convergence of $\mathcal{M}_2$ is its high parameter correlation (Cox and Reid, 1987), indicated by the parameter asymptotics in Section 4.3.1. The evidence of parameter correlation-induced slow convergence is that sharp decreases in the search space for $(\hat{\phi}, \hat{\alpha})$ do not lead to improved convergence rates. The convergence rate of $\mathcal{M}_1$ is superior in non-nugget effect models on the exponential Gaussian fields and identical to $\mathcal{M}_3$ on the Whittle Gaussian fields. Nonetheless, the convergence rate for $\mathcal{M}_1$ becomes more variable and exceeds that of $\mathcal{M}_3$ with nugget effects, which is further evidence of the vulnerability of $\mathcal{M}_1$ to nugget effects. Nugget effect models significantly decrease the convergence rate because there is one additional parameter τ^2 to estimate.

4.3. Increasing-domain asymptotics

We aim to investigate the asymptotic distributions of $\hat{\boldsymbol{\theta}}$. According to the asymptotic normality theory, we anticipate these estimates to follow a Gaussian distribution, with the covariance matrix being closely approximated by the inverse of the Fisher information matrix once the sample size n surpasses a certain threshold. Moreover, the off-diagonal elements of the inverse Fisher information

matrix indicate the explicit correlation between estimates. In their research, Stein (1999) examined the Fisher information of $\hat{\boldsymbol{\theta}}_2$ with sample sizes of 40 and 80 on both regular and irregular grids with various spacing factors.

The current experiment follows a similar procedure, but the contrast between regular and irregular grids is less noticeable due to the high density of observations ($n = 1600$). Irregular grids tend to create clusters of observations that provide more informative insights into the roles of (ν, β, ρ) in governing the spatial process' local behaviors (Stein, 1999). On the other hand, regular grids provide more information on α, consistent with the results in Stein (1999). Following Stein (1999), $\mathbf{I}_\phi$ is fixed as $n/(2\phi^2)$, where $\mathbf{I}_\phi$ denotes the corresponding diagonal element of the Fisher information matrix. Similarly, $\mathbf{I}_{\sigma^2}$ is also fixed across grid types and field smoothness at $n/(2\sigma^4)$. We observe that the Fisher information patterns do not vary with smoothness and nugget effects. Therefore, we provide Fisher information visualizations for the exponential Gaussian field on an irregular grid without nugget effects using 1600 samples. We use a spacing factor δ that varies from $(0.02, 0.05, 0.1, 0.2, 0.5, 1)$, multiplied by the two coordinates to obtain grids with different spacings. Moreover, we evaluate the uncertainty quantification's quality using the true asymptotic variance (TAV) and calculate the difference ratio (DRV) with respect to the sample variance (SV): DR $= |\text{SV} - \text{TAV}|/\text{TAV}$.

### 4.3.1. Asymptotic variances and correlations

This part of the experiment extends the asymptotic studies by Stein (1999) to a larger sample size in a two-dimensional field. Moreover, we also included $\mathcal{M}_1$ and $\mathcal{M}_3$ for the same analysis. We only used the first replicate for the analysis because the simulations reveal that the variations of $\mathbf{I}(\boldsymbol{\theta})$ are minimal for various replicates.

Figs. 8 and 9 illustrate that the MLEs of $\mathcal{M}_2$ distinguish themselves from those of $\mathcal{M}_1$ and $\mathcal{M}_3$, which have a high resemblance. Instead of taking up a continuously decreasing trend, such as $(\hat{\sigma}^2, \hat{\beta}, \hat{\rho}, \hat{\alpha})$, the TAV of $\hat{\phi}$ decreases first and then takes an upward trend on weak and medium fields, as presented in Fig. 8(d). Such upward-turning behavior results from strengthening the positive correlations of $(\hat{\phi}, \hat{\alpha})$ in δ (Fig. 9(d)), which also appeared in the work by Stein (1999). In addition, such behavior of $\hat{\phi}$ is not explicit on the strong field due to a weaker correlation of $(\hat{\phi}, \hat{\alpha})$, as depicted in Figs. 9(d). Moreover, $\hat{\alpha}$ exhibits opposite behavior to $(\hat{\beta}, \hat{\rho})$ in terms of the correlation with $\hat{\nu}$ displayed in Figs. 9(c), 9(f), and 9(i).

The TAVs of $(\hat{\phi}, \hat{\alpha})$ are also opposite to $(\hat{\sigma}^2, \hat{\beta}, \hat{\rho})$ concerning field strengths, which agrees with the patterns of their IQRs shown in Fig. 3. Diagonal elements of the Fisher information matrix for all these parameters at $\delta = 1$ are presented in Table 3 to address the differences together with the asymptotic correlations of the MLEs shown in Fig. 9.

Second, all parameters in $\mathcal{M}_2$ are positively correlated. However, in $\mathcal{M}_1$ and $\mathcal{M}_3$, the variance and range parameters are positively correlated (Figs. 9(a) and 9(g)), but the range and smoothness parameters are negatively correlated (Figs. 9(c) and 9(i)). In addition, their variance and smoothness parameters are close to being orthogonal (Figs. 9(b) and 9(h)). Although $\mathcal{M}_1$ and $\mathcal{M}_3$ display many similarities, some differences exist. A weaker negative correlation of $(\hat{\sigma}^2, \hat{\nu})$ and a stronger positive correlation of $(\hat{\sigma}^2, \hat{\rho})$ than their counterparts in $\mathcal{M}_1$ are displayed in Figs. 9(c), 9(i), 9(g), and 9(a). A stronger correlation of $(\hat{\beta}, \hat{\nu})$ is somewhat intuitive because β can be viewed as an aggregation of (ρ, ν), which in principle, should be more correlated with ν. In $\mathcal{M}_3$, such a correlation is diminished by taking $\nu^{1/2}$ out of the range parameter. There is no simple reason for a weaker correlation of $(\hat{\sigma}^2, \hat{\beta})$, but one possibility is that $\hat{\beta}$ contains $\hat{\nu}$, which is independent of $\hat{\sigma}^2$.

The TAVs of $\hat{\nu}$ rise in δ for all three parameterizations shown in Figs. 8(c), 8(f), and 8(i), which is consistent with the results from Stein (1999) because tightly spaced clusters are more informative for $\hat{\nu}$. In addition, the TAVs of $\hat{\nu}$ decrease when field correlation strengthens for all three parameterizations for various reasons. The decreases for $\mathcal{M}_1$ and $\mathcal{M}_3$ result from increase of $\mathbf{I}_\nu$ shown in Table 3 and also $\hat{\nu}$'s decreasing correlation with $\hat{\beta}$ and $\hat{\rho}$ shown in Figs. 9(c) and 9(i). However, for $\hat{\nu}$ in $\mathcal{M}_2$, the decrease of its TAV is due to its decreasing correlations with $(\hat{\phi}, \hat{\alpha})$ (Figs. 9(e) and 9(f)) because $\mathbf{I}_\nu$ decreases, as listed in Table 3. Decreasing TAVs of $\hat{\alpha}$ in δ agrees with the results from Stein (1999)'s experiment, which was explained by its relevance to low-frequency behaviors. Therefore, a sparse grid is more informative of $\hat{\alpha}$. The upward-turning behavior of the
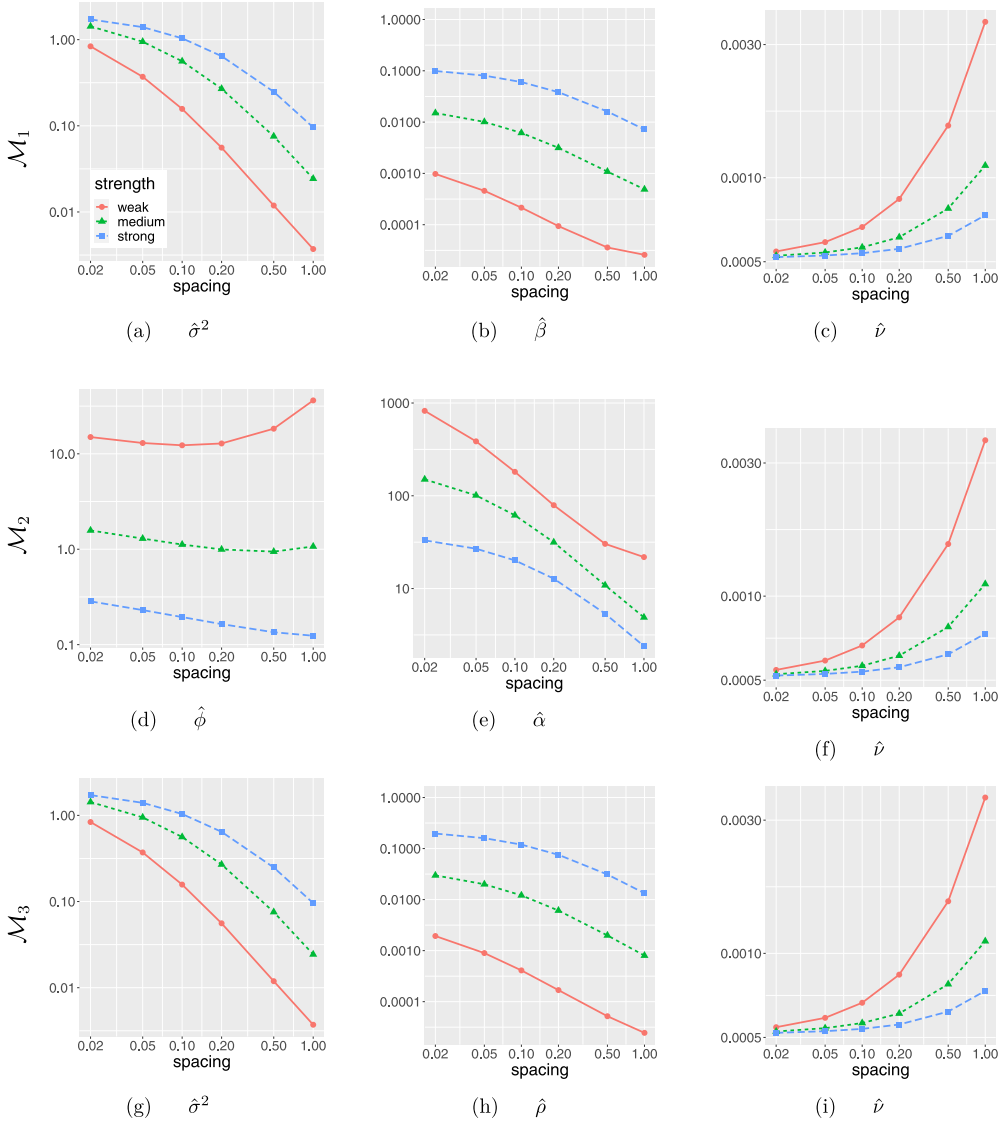
Fig. 8. True Asymptotic Variances (TAVs) for the MLEs of $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ under weak (red, circle), medium (green, triangle), and strong (blue, square) exponential Gaussian fields. The *x*-axis (spacing) and *y*-axis (TAV) except for $(\hat{\beta}, \hat{\rho})$ are on a logarithmic scale to better demonstrate the trend behaviors of the TAVs. The *y*-axis of $(\hat{\beta}, \hat{\rho})$ are on a linear scale because the decreasing trends are already obvious.

TAVs of $\hat{\alpha}$ in the work by Stein (1999) do not appear in the current studies because the asymptotic correlations of $(\hat{\phi}, \hat{\alpha})$ and $(\hat{\alpha}, \hat{\nu})$ are lower on irregular grids. Large TAVs of $(\hat{\beta}, \hat{\rho})$ at a small δ can be attributed to the failure to recognize the dependence range with observations squeezed into a small region and therefore, large uncertainties on the parameter estimates. The dependence range becomes easier to pinpoint as the study region expands. The TAV reductions of $\hat{\sigma}^2$ in δ are not intrinsic but rather due to $\hat{\sigma}^2$'s interactions with $(\hat{\beta}, \hat{\rho})$ because $\mathbf{I}_{\sigma^2}$ is fixed at 800 despite of varying δ and field correlation strengths. In addition, $(\hat{\sigma}^2, \hat{\nu})$ are nearly orthogonal as shown in Figs. 9(b) and
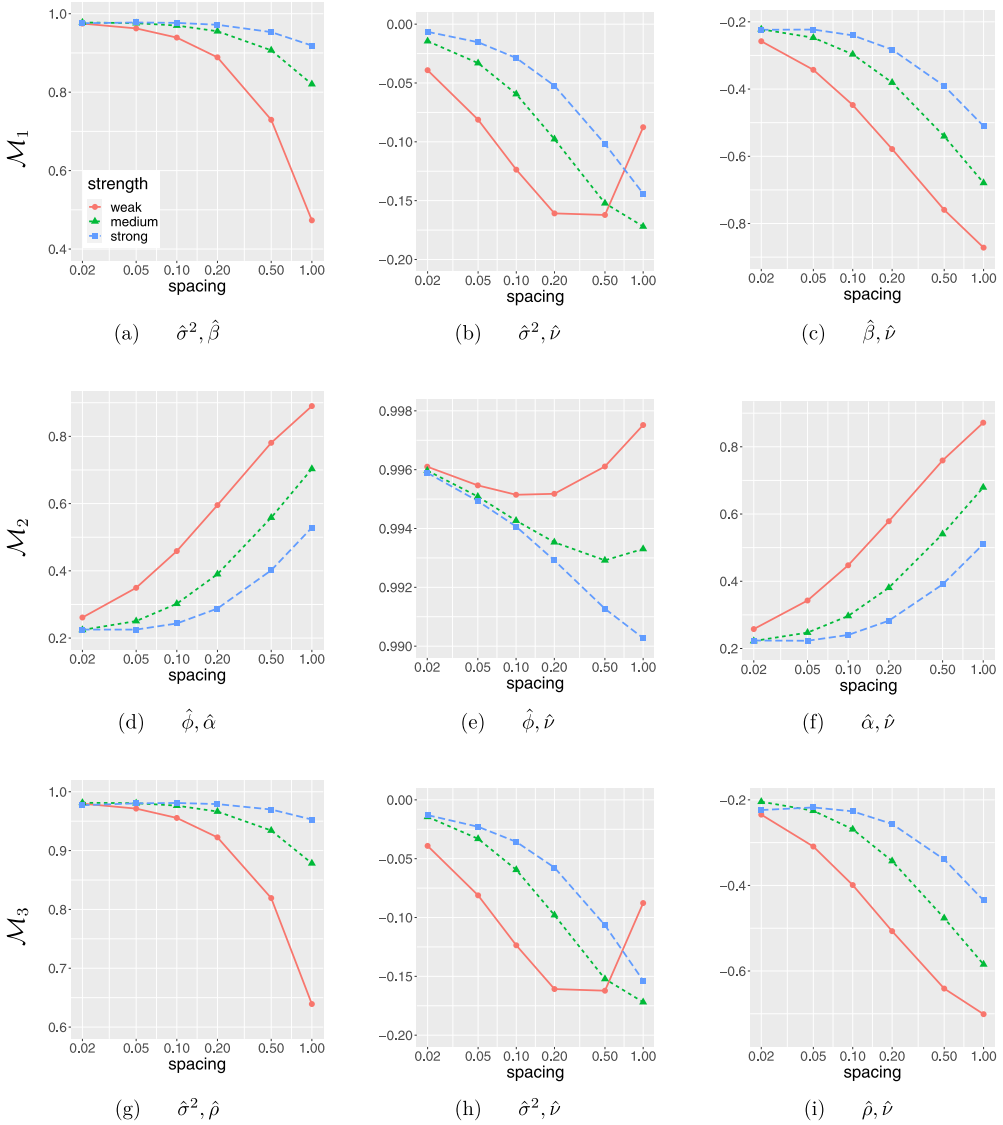
Fig. 9. True Asymptotic Correlations between MLEs of $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ under weak (red, circle), medium (green, triangle), and strong (blue, square) exponential Gaussian fields.

9(h). In particular, TAV of $\hat{\sigma}^2$ decreases in increasing δ primarily because of its decreasing correlation with $(\hat{\beta}, \hat{\rho})$ (Figs. 9(a) and 9(g)). Furthermore, TAVs of $\hat{\sigma}^2$ increase as field correlation strengthens, which can be ascribed to the rising correlations with $(\hat{\beta}, \hat{\rho})$.

As previously mentioned, Table 3 demonstrates that increase in field strength incurs tangible loss on $\mathbf{I}_\beta$ and $\mathbf{I}_\rho$. The correlation between range and variance parameters also increases in $\mathcal{M}_1$ and $\mathcal{M}_3$ (Figs. 9(a) and 9(g)). Consequently, TAVs of $(\hat{\beta}, \hat{\rho})$ tend to increase as field correlation strengthens. In contrast, $\mathbf{I}_\phi$ increases as field correlation strengthens. $\hat{\phi}$ also is less correlated with $(\hat{\alpha}, \hat{\nu})$ (Figs. 9(d) and 9(e)). Therefore, $\hat{\phi}$ exhibits opposite TAV behaviors compared with $(\hat{\beta}, \hat{\rho})$

Table 3

Fisher information for $\boldsymbol{\theta}$ for $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ on the exponential Gaussian fields ($\nu = 0.5$) at $\delta = 1$.

Dependence range/correlation strength ($\mathcal{M}_1$)	$\mathbf{I}_{\sigma^2}$	$\mathbf{I}_\beta$	$\mathbf{I}_\nu$
$\beta = 0.033$ (weak)	800.0	474 155.5	2653.6
$\beta = 0.1$ (medium)	800.0	71 786.2	10 694.5
$\beta = 0.234$ (strong)	800.0	14 146.3	21 987.3

Dependence range/correlation strength ($\mathcal{M}_1$)	$\mathbf{I}_\phi$	$\mathbf{I}_\alpha$	$\mathbf{I}_\nu$
$\phi = 9.6458$, $\alpha = 30.3030$ (weak)	8.6	0.3	74 835.8
$\phi = 3.1831$, $\alpha = 10.0000$ (medium)	79.0	0.4	71 747.9
$\phi = 1.3603$, $\alpha = 4.2735$ (strong)	432.3	0.6	71 265.5

Dependence range/correlation strength ($\mathcal{M}_1$)	$\mathbf{I}_{\sigma^2}$	$\mathbf{I}_\rho$	$\mathbf{I}_\nu$
$\rho = 0.0467$ (weak)	800.0	236 782.6	956.3
$\rho = 0.1414$ (medium)	800.0	35 917.6	6118.6
$\rho = 0.3309$ (strong)	800.0	7080.5	14 784.9

as indicated in Fig. 8. In addition, $\mathbf{I}_\alpha$ is only slightly increased and therefore, the TAV reductions originate from $\hat{\alpha}$'s decreasing correlations with $(\hat{\phi}, \hat{\nu})$ (Figs. 9(d) and 9(f)).

4.3.2. Quality of uncertainty quantification

This section calculates the SV, TAV, and DRV values to measure the quality of uncertainty quantification with TAV at finite samples, $n = 1600$. The computed results are summarized in Table S6 in the Supplementary Material. We have also repeated the experiment with different n and recorded the results in Tables S7, S8, and S9 to address the concerns of increasing n because when n is large enough, the variances should converge in theory regardless of parameterizations. We have shown that differences in matches of SV and TAV consistently persist as in $n = 1600$ with various values of n. Therefore, the comparisons drawn in this subsection are representative.

For the non-nugget effect models, Table S6 reveals that the DRs of $\hat{\nu}$ in $\mathcal{M}_2$ are the most significant in all settings, indicating apparent discrepancies between SV and TAV. Such discrepancies can be attributed to the strong parameter correlations (almost linear) of $(\hat{\phi}, \hat{\nu})$ as illustrated in Fig. 9(e). High parameter correlations critically affect MLEs, causing mathematical complexities and errors in their interpretations (Willmot, 1988). Difficulties will also arise in numerical optimizations, causing slow convergence or even divergence (Cox and Reid, 1987) and non-identifiability (Li and Vu, 2013). Consequently, the uncertainty for $\hat{\nu}$ in $\mathcal{M}_2$ is not properly approximated by the TAVs at finite samples.

In contrast, DRVs for $\hat{\nu}$ in $\mathcal{M}_1$ and $\mathcal{M}_3$ are small because $(\hat{\sigma}^2, \hat{\nu})$ are nearly independent, as depicted in Figs. 9(b) and 9(h). The strong correlation also affects $\hat{\phi}$, where the DRVs appear large in numerous settings. In addition, the DRs for $(\hat{\sigma}^2, \hat{\beta}, \hat{\rho})$ are notably larger when the variance and range estimates have almost linear dependence at medium and strong correlations. Although this disturbance is veiled by the inappropriate convergence tolerance for $\mathcal{M}_3$, the same issue arises when a lower tolerance is applied. Large DRs of $\hat{\alpha}$ are more likely to result from low Fisher information and, therefore, are challenging to estimate because $\hat{\alpha}$ does not display linear dependence with other parameters like $(\hat{\beta}, \hat{\rho})$. Furthermore, significant DRVs for $(\hat{\sigma}^2, \hat{\rho})$ confirm their high-tolerance induced bias shown in Fig. 3. Overall, $\mathcal{M}_1$ achieves the best match between SVs and TAVs in the non-nugget effect models.

In addition, the TAV of $\hat{\sigma}^2$ and $\hat{\nu}$ are comparable in both $\mathcal{M}_1$ and $\mathcal{M}_3$ but the TAVs of $\hat{\beta}$ are obviously smaller than those of $\hat{\rho}$ in most cases. An appropriate explanation of the smaller TAVs of $\hat{\beta}$ can also be drawn from the link function, $\beta = 2\nu^{1/2}\rho$, provided in Table 1. If we assume ν as known and apply the invariance property of MLEs, we have $\hat{\beta} = 2\nu^{1/2}\hat{\rho}$. Then, $\text{Var}(\hat{\rho}) = 4\nu\text{Var}(\hat{\beta})$. In principle, $\text{Var}(\hat{\rho})$ is expected to exceed $\text{Var}(\hat{\beta})$ except on extremely rough fields (i.e., $\nu < 0.25$). The situation appears on the weak exponential Gaussian field but not on the others, as shown in Table S6.

Nugget effects increase the TAVs and result in considerable DRs for $\hat{\boldsymbol{\theta}}_1$, $\hat{\boldsymbol{\theta}}_2$, which is consistent with the heavier effects of τ^2 on $\mathcal{M}_1$ and $\mathcal{M}_2$ discussed in Section 4.2.1. In numerous settings, TAVs

of $\hat{\beta}$ can exceed those of $\hat{\rho}$. DRVs of $\hat{\theta}_3$ are relatively contained. Moreover, the nugget effects are more critical on strongly high-frequency relevant parameters. For instance, TAVs predominantly scale by ten times or more for $(\hat{v}, \hat{\phi})$. However, TAV increases are moderate on the rest, primarily scaling by 4 to 5 times at maximum. The TAVs for $\hat{\beta}$ tend to scale larger than those for $\hat{\rho}$, again indicating the susceptibility of $\mathcal{M}_1$ to the nugget effects compared with $\mathcal{M}_3$. Furthermore, Table S5 in the Supplementary Material indicates that $\mathcal{M}_3$ offers the most stable DRVs for $\hat{\tau}^2$, whereas $\mathcal{M}_1$ and $\mathcal{M}_2$ display critical deviations in some cases. Therefore, TAVs of $\hat{\theta}_3$ approximate the uncertainties more efficiently at finite samples in the nugget effect models. The TAVs of $\hat{\tau}^2$ follow the same pattern as the IQRs shown in Fig. 4. The TAV of $\hat{\tau}^2$ reduces as the field correlation and smoothness increase and drops to the lowest on a strong-correlated smooth field. It is difficult to draw conclusions on the asymptotic distributions of the MLEs because 1600 samples might not be sufficient for the MLEs to reach normality. The number of samples required for asymptotic normality requires further explorations.

5. Tile low-rank approximation

The direct maximization of the log-likelihood function (4) is usually time-consuming and computationally inefficient. Abdulah et al. (2018b) implemented a tile-based parallel TLR approximation method for the MLE operation to compress off-diagonal tiles up to a pre-specified accuracy. First, the TLR decomposes the large covariance matrix into tiles of a given tile size, $\Sigma(\theta) = \{\mathbf{D}(\theta)_{i,j}\}$, and maintains the diagonal tiles $\mathbf{D}(\theta)_{i,i}$ in dense forms. Singular value decomposition is performed on the individual off-diagonal tiles so that only a certain number of eigenvalues and eigenvectors can be preserved to reach the specified accuracy measured by the Frobenius norm, $\|\mathbf{D}(\theta)_{i,j} - \tilde{\mathbf{D}}(\theta)_{i,j}\|_F < \epsilon$. The low-rank tile $\tilde{\mathbf{D}}(\theta)_{i,j}$ occupies less memory space during computation, accelerating matrix operations. Abdulah et al. (2018b) demonstrated that the TLR method could significantly reduce modeling time while maintaining decent estimation accuracy. Furthermore, Hong et al. (2021) argued that TLR outperforms the Gaussian predictive process proposed by Banerjee et al. (2008) and the composite likelihood method developed by Vecchia (1988) and Curriero and Lele (1999).

In this section, we evaluate the performance of the TLR approximation on the previously generated datasets from the three parameterizations under sparse covariance settings. We set the tile size to 400 and the accuracy level to 10^{-7}. Therefore, the covariance matrix was divided into 16 tiles, each containing 400^2 entries. We applied the TLR approximation to the 12 off-diagonal tiles while keeping the diagonal tiles dense. As reported in Section 4.2.1, the measures MSPE and MLOE are insensitive to the parameterizations. Hence, we only present the DRs between the sample means and true values, the AER, and MMOM on the exponential Gaussian fields (similar results were observed on Whittle Gaussian fields). Overall, the TLR method resembles the results produced by the exact method. In addition, $\mathcal{M}_1$ is slightly more accurate than $\mathcal{M}_3$ for the non-nugget effect models in terms of point estimations (Table S2, Supplementary Material) and AERs (Table S4, Supplementary Material) but is more vulnerable to the nugget effects. The MMOMs also reflect the same situation: the MMOMs are mostly comparable for the non-nugget effect models, whereas the MMOMs of $\mathcal{M}_1$ tend to deviate more from zero compared with MMOMs of $\mathcal{M}_3$ in the nugget effect models shown in Fig. 10, indicating that using $\mathcal{M}_1$ in TLR settings is more prone to model specification errors.

Another crucial difference in parameterizations is that applying $\mathcal{M}_2$ in TLR settings is quite challenging. The problem is that ϕ and α cannot land at small values in the search space because they will then characterize fields of significantly strong correlations, for which the resulting covariance matrix has a high possibility of being numerically singular if a loose accuracy level is applied to the off-diagonal blocks. Indeed, we encountered this problem when the accuracy level was 10^{-5}, 10^{-7}, and 10^{-9}.

The results from Abdulah et al. (2018b) demonstrated that a low accuracy level increases the modeling time and diminishes the purpose of the complexity reduction. Numerical singularity is not a critical issue for $\mathcal{M}_1$ and $\mathcal{M}_3$ because small range parameter values typically characterize weakly correlated fields. Although numerical singularity is not a significant concern, using the nugget effect models in TLR settings is preferred. Noise added to the diagonals can improve the matrix condition number and, thus, better avoid numerical singularity, enabling even looser accuracy conditions and further accelerating the modeling process.
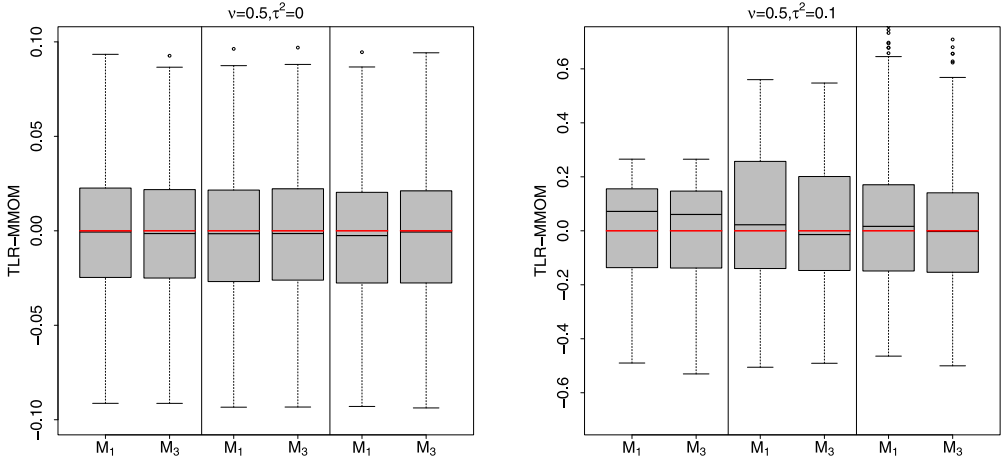
Fig. 10. Boxplots of the MMOM using TLR approximation with 300 replicates for the non-nugget effect models ($\tau^2 = 0$) and nugget effect models ($\tau^2 = 0.1$) under weak (first block), medium (second block), and strong (third block) field correlations, separated by vertical lines in each sub-figure, on the exponential Gaussian fields ($\nu = 0.5$). Red lines denote the expected MMOM value when the true covariance is applied. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
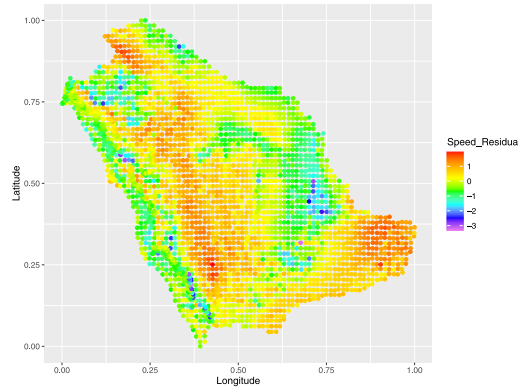
Table 4
Comparisons of convergent iterations of the MLEs for $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ using Saudi wind speed residual data under exact covariance estimation with and without nugget effects.

Model/parameterization	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
Non-nugget	187.4 (64.5)	314.7 (85.2)	189.7 (99.9)
Nugget	308.9 (168.2)	545.1 (224.1)	296.3 (101.0)

6. Application to Saudi wind speed data

In this section, we apply $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ to Saudi Arabia wind speed residual data from Huang et al. (2022) visualized in Fig. 11, containing 3173 observations across the entire country on January 1st, 2013. The locations are rescaled within $[0, 1]^2$ to match the previous experiment settings. In addition, we randomly split the data into 10 sets of training and testing data, accounting for 80% (2538) and 20% (635) of the entire dataset. Then, we measured the convergent iterations and asymptotic variance (AV) estimated from the testing data and MSPE on the training data. The tuning parameters are the same as in the simulation studies. The numbers in Tables 4 to 6 indicate the means computed across all 10 testing and training sets. The numbers in parentheses indicate their associated standard deviations. Tables 4 to 6 demonstrate that many of the conclusions from numerical simulations remain true in practice. First, we observe that $\mathcal{M}_2$ requires the most convergent iterations for both non-nugget and nugget effect models. In addition, $\mathcal{M}_1$ is the most efficient for the non-nugget effect models but becomes volatile and gets outperformed by $\mathcal{M}_3$ in the presence of nugget effects. Further, $(\hat{\sigma}^2, \hat{\rho})$ in $\mathcal{M}_3$ encounter inaccuracy issues in the non-nugget model, as the AVs significantly deviate from those of their counterparts in $\mathcal{M}_1$ with the nugget effect. The nugget effect τ^2 and its AV are consistently estimated to be around 0.03 and 0.00012 across all parameterizations. Therefore, the overall nugget effect is limited. Hence, the AVs of $(\hat{\sigma}^2, \hat{\rho})$ should be similar in both scenarios like those of $(\hat{\sigma}^2, \hat{\beta})$ in $\mathcal{M}_1$.

Second, $\hat{\alpha}$ in both scenarios and $\hat{\phi}$ in the nugget effect model display much more significant AVs than the other estimates. Third, MSPE barely changes across differing parameterizations and increases under the exposure of nugget effects. In addition, the nugget effect, even minimally, significantly affects $(\hat{\phi}, \hat{\nu})$ but has limited influences on the uncertainties of $(\hat{\sigma}^2, \hat{\beta}, \hat{\rho}, \hat{\alpha})$ in comparison.

Fig. 11. Saudi wind speed residual data on January 1st, 2013, at 3173 locations.

Table 5
Comparisons of the MSPE for $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ using Saudi wind speed residual data under an exact covariance estimation with and without nugget effects.

Model/parameterization	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
Non-nugget	0.1115 (0.0166)	0.1089 (0.0142)	0.1076 (0.0156)
Nugget	0.1121 (0.0163)	0.1118 (0.0166)	0.1107 (0.0156)

Table 6
Comparisons of AVs for $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ using Saudi wind speed residual data under an exact covariance estimation with and without nugget effects.

$\mathcal{M}_1$	σ^2	β	ν	τ^2
	0.1040 (0.1450)	0.0546 (0.0873)	0.000258 (0.0000215)	0
$\mathcal{M}_2$	ϕ	α	ν	τ^2
	0.00649 (0.00221)	5.7512 (0.9247)	0.000281 (0.0000222)	0
$\mathcal{M}_3$	σ^2	ρ	ν	τ^2
	0.5272 (0.7103)	0.5897 (0.8085)	0.000249 (0.0000233)	0
$\mathcal{M}_1$	σ^2	β	ν	τ^2
	0.1671 (0.3272)	0.0413 (0.0892)	0.00423 (0.00157)	0.000123 (0.0000326)
$\mathcal{M}_2$	ϕ	α	ν	τ^2
	3.3101 (5.6246)	11.1619 (2.2612)	0.00557 (0.00157)	0.000119 (0.0000346)
$\mathcal{M}_3$	σ^2	ρ	ν	τ^2
	0.0683 (0.1209)	0.0304 (0.0647)	0.00478 (0.00154)	0.000127 (0.0000440)

Last, the AV of $\hat{\beta}$ is slightly more significant but roughly identical to that of $\hat{\rho}$ in the nugget effect model because $\hat{\nu}$ is estimated to be around 0.3, which is a rough field with some minor nugget effects.

7. Discussion

In this study, we comprehensively compared the three most widely used parameterizations of the Matérn covariance function. We utilized both simulated and real data to investigate various aspects such as modeling efficiency and accuracy, the behavior of maximum likelihood estimators, and other crucial factors. Our analysis was made possible by utilizing the high-performance computing software, *ExaGeoStat*, which allowed us to perform matrix operations in parallel and

generate large-scale spatial data efficiently. We also employed benchmarking functions such as MMOM and the Fisher information matrix, which are sensitive to different parameterizations and can aid in selecting the optimal parameterization.

In summary, $\mathcal{M}_1$ is preferable over $\mathcal{M}_2$ and $\mathcal{M}_3$ in modeling speed. Further, $\mathcal{M}_1$ also provides decent and stable estimation accuracy in non-nugget effect models, offering lower uncertainty than $\mathcal{M}_2$ and avoiding the high-tolerance inaccuracy issues occurring in $\mathcal{M}_3$. However, in the presence of nugget effects, $\mathcal{M}_1$ and $\mathcal{M}_2$ are more disturbed than $\mathcal{M}_3$ with biased estimates and apparent deviations between TAVs and SVs. Moreover, the resistance of $\mathcal{M}_3$ to the nugget effects also holds in TLR approximations. The TAV of MLEs for $\mathcal{M}_1$ is highly likely to be the lowest in non-nugget effect models, allowing for more informative confidence intervals. Prediction efficiency and the MSPE of the three parameterizations are comparable, but the MSPEs of $\mathcal{M}_1$ and $\mathcal{M}_3$ are the most trustworthy of the three. Moreover, $\mathcal{M}_1$ and $\mathcal{M}_3$ offer better parameter orthogonality, defined in Cox and Reid (1987), than $\mathcal{M}_2$; Jørgensen and Knudsen (2004) also refer to this concept of parameter orthogonality as F-orthogonality, indicating the asymptotic independence among the MLEs. Therefore, TAVs of $\hat{\nu}$ are closer to the SVs. Furthermore, parameters for $\mathcal{M}_1$ and $\mathcal{M}_3$ have more statistical-intuitive interpretations, in which $\sigma^2 + \tau^2 = \text{Var}\{Z(\mathbf{s})\}$ and β and ρ are direct measures of the dependence range. In addition, the TLR method can be applied to $\mathcal{M}_1$ and $\mathcal{M}_3$, whereas $\mathcal{M}_2$ often runs into numerical singularity issues. Lastly, per the small variances of MLEs for $\mathcal{M}_1$ and $\mathcal{M}_3$, it would be relatively safe to use a smaller search space in optimization, further enhancing modeling speed. However, such action is hazardous for $\mathcal{M}_2$, whose MLEs might take a large range of values and vary dramatically. Due to the above reasoning, we recommend $\mathcal{M}_1$ for the non-nugget effect model and $\mathcal{M}_3$ for the nugget effect model. If there is an absolute preference, the link functions presented in Table 1 and the Jacobian transformations of $\mathbf{I}(\boldsymbol{\theta})$ provided in the Supplementary Material can enable flexible transitions between any of them. As for the Bayesian context, we believe similar issues as in the frequentist context will likely arise due to the likelihood part unless very strong priors are imposed. As for the Bayesian context, we believe that similar issues as in the frequentist context will likely arise due to the likelihood part unless very strong priors are imposed. Further analysis of the Matérn parametrizations in the Bayesian context is worth exploring but beyond the scope of the current paper.

Acknowledgments

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spasta.2023.100787.

References

Abdulah, S., Cao, Q., Pei, Y., Bosilca, G., Dongarra, J., Genton, M.G., Keyes, D.E., Ltaief, H., Sun, Y., 2021. Accelerating geostatistical modeling and prediction with mixed-precision computations: A high-productivity approach with PaRSEC. IEEE Trans. Parallel Distrib. Syst. 33 (4), 964–976.

Abdulah, S., Li, Y., Cao, J., Ltaief, H., Keyes, D.E., Genton, M.G., Sun, Y., 2023. Large-scale environmental data science with ExaGeoStatR. Environmetrics 34 (1), e2770.

Abdulah, S., Ltaief, H., Sun, Y., Genton, M.G., Keyes, D.E., 2018a. ExaGeoStat: A high performance unified software for geostatistics on manycore systems. IEEE Trans. Parallel Distrib. Syst. 29 (12), 2771–2784.

Abdulah, S., Ltaief, H., Sun, Y., Genton, M.G., Keyes, D.E., 2018b. Parallel approximation of the maximum likelihood estimation for the prediction of large-scale geostatistics simulations. In: 2018 IEEE International Conference on Cluster Computing (CLUSTER). pp. 98–108.

Abt, M., Welch, W.J., 1998. Fisher information and maximum-likelihood estimation of covariance parameters in Gaussian stochastic processes. Canad. J. Statist. 26 (1), 127–137.

Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H., 2008. Gaussian predictive process models for large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (4), 825–848.

Cao, Q., Abdulah, S., Alomairy, R., Pei, Y., Nag, P., Bosilca, G., Dongarra, J., Genton, M.G., Keyes, D.E., Ltaief, H., et al., 2022. Reshaping geostatistical modeling and prediction for extreme-scale environmental applications. In: 2022 SC22: International Conference for High Performance Computing, Networking, Storage and Analysis (SC). IEEE Computer Society, pp. 13–24.

Cox, D.R., Reid, N., 1987. Parameter orthogonality and approximate conditional inference. J. R. Stat. Soc. Ser. B Stat. Methodol. 49 (1), 1–18.

Curriero, F.C., Lele, S., 1999. A composite likelihood approach to semivariogram estimation. J. Agric. Biol. Environ. Stat. 4 (1), 9–28.

De Oliveira, V., Han, Z., 2022. On information about covariance parameters in Gaussian Matérn random fields. J. Agric. Biol. Environ. Stat. 27 (4), 690–712.

Geoga, C.J., Anitescu, M., Stein, M.L., 2020. Scalable Gaussian process computations using hierarchical matrices. J. Comput. Graph. Statist. 29 (2), 227–237.

Geoga, C.J., Marin, O., Schanen, M., Stein, M.L., 2023. Fitting Matérn smoothness parameters using automatic differentiation. Stat. Comput. 33 (2), 48.

González-Santander, J., 2018. Closed-form expressions for derivatives of Bessel functions with respect to the order. J. Math. Anal. Appl. 466 (1), 1060–1081.

Gough, B., 2009. GNU Scientific Library Reference Manual. Network Theory Ltd..

Handcock, M.S., Stein, M.L., 1993. A Bayesian analysis of kriging. Technometrics 35 (4), 403–410.

Handcock, M.S., Wallis, J.R., 1994. An approach to statistical spatial-temporal modeling of meteorological fields. J. Amer. Statist. Assoc. 89 (426), 368–378.

Hong, Y., Abdulah, S., Genton, M.G., Sun, Y., 2021. Efficiency assessment of approximated spatial predictions for large datasets. Spatial Stat. 43, 100517.

Huang, H., Castruccio, S., Genton, M.G., 2022. Forecasting high-frequency spatio-temporal wind power with dimensionally reduced echo state networks. J. R. Stat. Soc. Ser. C Appl. Stat. 71 (2), 449–466.

Jørgensen, B., Knudsen, S.J., 2004. Parameter orthogonality and bias adjustment for estimating functions. Scand. J. Stat. 31 (1), 93–114.

Li, P., Vu, Q.D., 2013. Identification of parameter correlations for parameter estimation in dynamic biological models. BMC Syst. Biol. 7 (1), 1–12.

Loh, W.-L., 2005. Fixed-domain asymptotics for a subclass of Matérn type Gaussian random fields. Ann. Statist. 33 (5), 2344–2394.

Mardia, K.V., Marshall, R.J., 1984. Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika 71 (1), 135–146.

Matérn, B., 1960. Spatial Variation: Stochastic Models and Their Application to Some Problems in Forest Surveys and Other Sampling Investigations, Vol. 49. Medd. Statens Skogsforskningsinstitut, Stockholm, (5).

Myung, J.I., Navarro, D.J., et al., 2005. Information matrix. In: Encyclopedia of Statistics in Behavioral Science, Vol. 2. Citeseer, pp. 923–924.

Nychka, D., Furrer, R., Paige, J., Sain, S., 2021. Fields: Tools for spatial data. R package version 14.1.

Olver, F.W., Lozier, D.W., Boisvert, R.F., Clark, C.W., 2010. NIST Handbook of Mathematical Functions Hardback and CD-ROM. Cambridge University Press.

Petersen, K.B., Pedersen, M.S., et al., 2008. The matrix cookbook. Tech. Univ. Den. 7 (15), 510.

Porcu, E., Bevilacqua, M., Schaback, R., Oates, C.J., 2023. The Matérn model: A journey through statistics, numerical analysis and machine learning. arXiv preprint arXiv:2303.02759.

Powell, M.J., 2009. The BOBYQA Algorithm for Bound Constrained Optimization Without Derivatives, Vol. 26. Cambridge NA Report NA2009/06, University of Cambridge, Cambridge.

Ribeiro, P.J., Diggle, P.J., 2001. GeoR: a package for geostatistical analysis. R-NEWS (ISSN: 1609-3631) 1 (2), 14–18.

Schlather, M., Malinowski, A., Oesting, M., Boecker, D., Strokorb, K., Engelke, S., Martini, J., Ballani, F., Moreva, O., Auel, J., Menck, P.J., Gross, S., Ober, U., Ribeiro, P., Singleton, R., Pfaff, B., R Core Team, 2019. RandomFields: Simulation and analysis of random fields. R package version 3.3.1.

Stein, M.L., 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer Science & Business Media.

Stein, M.L., 2014. Limitations on low rank approximations for covariance matrices of spatial data. Spatial Stat. 8, 1–19.

Vecchia, A.V., 1988. Estimation and model identification for continuous spatial processes. J. R. Stat. Soc. Ser. B Stat. Methodol. 50 (2), 297–312.

Willmot, G.E., 1988. Parameter orthogonality for a family of discrete distributions. J. Amer. Statist. Assoc. 83 (402), 517–521.

Zimmerman, D.L., Cressie, N., 1992. Mean squared prediction error in the spatial linear model with estimated covariance parameters. Ann. Inst. Statist. Math. 44 (1), 27–43.