Computational Statistics and Data Analysis 173 (2022) 107503

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

Bayesian multiresolution modeling of georeferenced data: An extension of 'LatticeKrig'

John Paige [a,*], Geir-Arne Fuglstad [a], Andrea Riebler [a], Jon Wakefield [b]

[a] Department of Mathematical Sciences, NTNU, Trondheim, Norway
[b] Department of Statistics and Biostatistics, University of Washington, Seattle, USA

ARTICLE INFO

Article history:
Received 2 February 2021
Received in revised form 8 April 2022
Accepted 8 April 2022
Available online 21 April 2022

MSC:
62H11
62P99

Keywords:
Spatial analysis
Extended LatticeKrig
Latent Gaussian models
Bayesian inference
Integrated nested Laplace approximations

ABSTRACT

'LatticeKrig' (LK) is a spatial model that is often used for modeling multiresolution spatial data with flexible covariance structures. An extension to LK under a Bayesian framework is proposed that uses integrated nested Laplace approximations (INLA). The extension enables the spatial analysis of non-Gaussian responses in latent Gaussian models, joint spatial modeling with structured and unstructured random effects, and native support for multithreaded parallel likelihood computation. The proposed extended LatticeKrig (ELK) model uses a reparameterization of LK so that the parameters and prior selection are intuitive and interpretable. Priors can be used to make inference robust by penalizing more complex models, and integration over model parameters allows for posterior uncertainty estimates that account for uncertainty in covariance parameters. ELK's ability to reliably resolve multiresolution spatial structure for pointwise and areal predictions is demonstrated in both simulation study and two applications with non-Gaussian observations: a set of 188,717 LiDAR forest canopy height observations in Bonanza Creek Experimental Forest in Alaska, and a set of 1,612 clusters containing counts of secondary education completion from the 2014 Kenya demographic health survey. ELK has improved central predictions as well as uncertainty characterization according to the considered scoring rules when compared against a number of other models, particularly in the forest canopy height application, and performed faster than LK in our tests in part due to its support for and use of parallelization.

1. Introduction

The increasing size and complexity of spatial point datasets in fields such as climatology, public health, ecology, and the social sciences have been concurrent with methodological developments in spatial statistics. One popular method of spatial analysis developed to handle such datasets is known as 'LatticeKrig' (LK) (Nychka et al., 2015), with an associated package sharing the name (Nychka et al., 2016). Unlike many other models that are able to handle large spatial datasets, LK focuses on the ability to model complex spatial dependence structures, particularly those that exhibit multiple scales of spatial correlation, by separating the modeled spatial process into multiple 'layers', each with its own spatial resolution and correlation scale.

* Corresponding author.
E-mail address: john.paige@ntnu.no (J. Paige).

https://doi.org/10.1016/j.csda.2022.107503

While there are a number of other available methods that can, in theory, be used to model large datasets exhibiting multiscale spatial behavior, such as Vecchia approximations (Katzfuss and Guinness, 2020; Katzfuss et al., 2018; Zilber and Katzfuss, 2019), and related methods, their available implementations provide no simple way of accounting for this behavior in practice. For a review of general spatial methods for 'big data', see Heaton et al. (2019). Autoregressive co-kriging, introduced in Kennedy and O'Hagan (2000), is able to account for multiple datasets, each observed at different resolutions. However, it is not intended for modeling multiscale spatial correlations in a single dataset.

Although LK is a very useful tool for modeling multiscale spatial data, it has a number of shortcomings that we hope to address in this work. LK is unable to model non-Gaussian responses, which considerably limits its applicability in practice. Moreover, the `LatticeKrig` package is unable to account for multiscale spatial correlation jointly with structured or unstructured random effects. Many nonlinear covariate or temporal effect models, such as those involving random walks or some kinds of smoothing splines, incorporate such random effects, and would therefore be unable to be included within the `LatticeKrig` package. While LK allows for more flexible covariance modeling than is typically allowed under other models assuming isotropic spatial covariances, `LatticeKrig` does not allow for any way to penalize this added flexibility or encourage simpler models, and little guidance is given as to how to structure the spatial layers and how to choose their resolutions. Lastly, `LatticeKrig` does not natively support multithreaded parallel computations, which limits its applicability to large datasets.

These limitations prevent LK from being applied to datasets that might otherwise be ideally suited for it due to their multiresolution spatial structure. In LiDAR data, for instance, observations are collected along long, thin tracts and in potentially very high resolution relative to the size of the spatial domain (Abdalati et al., 2010; Schutz et al., 2005; Dubayah et al., 2020; Cook et al., 2013). Such data can be used to identify fine scale spatial structure if present, but also contain gaps between the measurement transects where long scale correlations may be especially important for prediction. Finley et al. (2020b) consider a forest canopy height dataset with LiDAR measurements, and find evidence suggesting that forest canopy height is associated nonlinearly with percent tree cover. Since forest canopy height is nonnegative, it is also inherently non-Gaussian. Despite the potential for multiscale spatial structure, LK is not well suited to analyze this dataset due to the non-Gaussian responses as well as the nonlinear effect of percent tree cover.

In other cases, where areally aggregated predictions are required, a model's ability to accurately account for multiple correlation scales can be especially important. In Appendix A, we show that uncertainties in an areal average are most impacted by correlation ranges close to the radius of that area. Different correlation scales therefore become important when producing areal averages over different sized areas. Estimates of population averages require such areal averaging, and are of great importance in survey statistics and small area estimation, where estimates of population averages may be desired over multiple different administrative areas with different shapes and sizes. Many population quantities of interest in such contexts are based on non-Gaussian responses, such as population prevalence estimates that typically rely on count data. Estimating the prevalence of women's secondary education completion, for instance, relies on population count data, and is important for Sustainable Development Goal (SDG) 4 (United Nations, 2020), which calls for improvements in secondary education to the point where everyone can complete their secondary education by 2030 regardless of their gender or where they live.

Although multiscale spatial correlations may be relevant in many applications, the most common approach to spatial modeling is to use parametric classes of spatial covariance functions such as the Matérn family that do not necessarily account for multiscale structure. It is known that, under infill asymptotics, it is the behavior of the Matérn covariance function at short spatial scales that most affects the likelihood and pointwise predictions (Stein, 1999, Ch. 3). This means that while short scale behavior of the Matérm covariance may be fit accurately, long scale correlations in the data will often not be accurately reproduced by the fit model. However, as we later show in the simulation study, long range correlations become increasingly important when making predictions far from observations, and for certain areal averages.

The difficulty in identifying spatial covariance parameters can make it especially important to integrate over their uncertainty when estimating uncertainty in the predictions, and when an estimate of the covariance function is desired. In a frequentist setting, and in LK, the bootstrap can be applied, but it relies on asymptotics, and is computationally expensive since it requires the model to be refit many times (Sjöstedt-de Luna and Young, 2003). While conditional simulation (see, e.g. Lantuéjoul 2013) is a computationally tractable frequentist method for estimating predictive uncertainty, it does not account for uncertainty in the covariance parameters. Handock and Stein (1993) and Gelfand et al. (2010, Ch. 3.7) recommend using Bayesian inference in spatial statistics due to the importance of accounting for uncertain covariance structure when making predictions and also note that prior information can improve model identifiability when available. However, Markov Chain Monte Carlo (MCMC) techniques are often difficult to implement with long running times and large memory requirements, especially with large numbers of observations (Filippone et al., 2013). Detailed output diagnostics are also necessary to assess convergence.

As such, the key limitation in providing Bayesian inference for multiresolution spatial models is the computational complexity involved. In this work, we propose to take advantage of the deterministic algorithm for Bayesian inference based on Integrated Nested Laplace Approximations (INLA) (Rue et al., 2009). INLA is able to account for uncertainty in model parameters by approximating the joint posterior distribution of the model hyperparameters using a multivariate spline. INLA then samples from the joint hyperparameter distribution, and, conditional on each sample from the hyperparameter posterior, INLA samples from the rest of the posterior. Rather than conditioning on a single set of hyperparameter estimates, this allows for INLA to account for hyperparameter uncertainty when generating a predictive distribution. LK uses different layers

of compact basis functions together with an associated sparse precision matrix, and we show how it can be included in the INLA framework of latent Gaussian models. We provide an implementation using the `R` package `INLA`, which permits fast and accurate estimation of posterior marginal densities provided that the number of parameters is not too big (typically 2 to 5, but not exceeding 20 (Rue et al., 2017)). This extended version of LK is termed extended LatticeKrig (ELK). A key change from the original LK formulation is a reparametrization that improves interpretability and facilitates modeling and prior selection. Furthermore, the INLA implementation allows the ELK spatial model to be fit jointly with other random effects such as models for temporal trends or nonlinear covariate effects, to handle non-Gaussian responses, to integrate over parameter uncertainty, and to incorporate prior knowledge through expert knowledge and/or for the purpose of robustness. Importantly for the purposes of analyzing large datasets, ELK's implementation in INLA gives it native support for multithreaded parallel computation, unlike LK.

We will contrast ELK to different spatial models throughout this work in order to better understand its performance relative to existing methods. A number of state of the art spatial models are reviewed in Heaton et al. (2019), although in this work we focus in particular on comparing ELK to LK when possible.

In Section 2 we introduce the two main applications that motivated this work, namely forest canopy height in Bonanza Creek Experimental Forest and secondary education prevalence for women in Kenya. In Section 3 we describe LK and ELK. We evaluate ELK, LK, and a stochastic partial differential equation (SPDE) Matérn model in a simulation scenario when fit to random fields with realistic mixtures of short and long-range correlations in Section 4, demonstrating the importance of multiples scales of spatial correlation in both point and areal predictions far from the nearest observation. In Sections 5 and 6 the ELK model is applied to real data introduced in Section 2 where mutliresolution spatial structure is relevant for prediction, and ELK's predictive performance is assessed relative to LK when possible, and an SPDE model otherwise. Section 7 concludes this work with a discussion.

2. Motivating applications

2.1. Forest canopy height in Bonanza Creek experimental forest

A forest can influence nearby ecosystems in many ways, and forest structure is often a key input to ecosystem and earth system models (Klein et al., 2015; Hurtt et al., 2004; Finney, 1998). There is not only a demand for national and biome scale forest canopy height (FCH) estimates, but also an increasing ability to collect high quality and large scale LiDAR data. Examples of LiDAR systems capable of such mapping efforts include ICESat-2 (Abdalati et al., 2010; Schutz et al., 2005), Global Ecosystem Dynamics Investigation LiDAR (Dubayah et al., 2020), and NASA Goddard's LiDAR, Hyperspectral, and Thermal (G-LiHT) Airborne Imager (Cook et al., 2013).

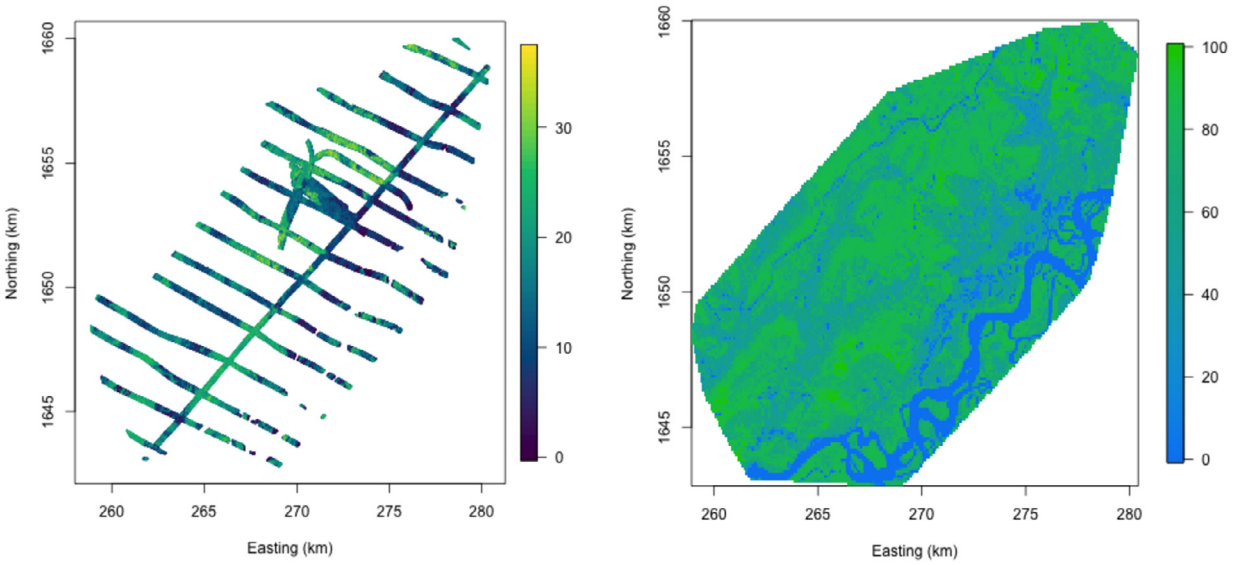
In this application, revisited in Section 5, we will consider a G-LiHT LiDAR forest canopy height dataset in Bonanza Creek Experimental Forest (BCEF) in Alaska, USA. The dataset consists of 188,717 observations. Our goal is to both smooth estimates of FCH over the support of the data if necessary, and in particular to provide estimates of FCH between the LiDAR data transects. Since FCH is nonnegative, Gaussian response models will likely be problematic. We will also use Landsat derived estimates of percent tree cover (PTC) from Hansen et al. (2013) as a covariate in the prediction. Since past evidence suggests the association between FCH and PTC may be nonlinear (Finley et al., 2020b), we will model this nonlinear association with ELK using a structured random effect model specified in Section 5. Both FCH and PTC are shown in Fig. 1. The dataset is available in the `spNNGP` R package (Finley et al., 2020a).

ELK may be relevant for this particular application due to its ability to model both long and short scale spatial structure, to handle large datasets of non-Gaussian responses, and to account for the nonlinear effect of PTC on FCH using structured random effect models jointly with the spatial effect. In this dataset, the LiDAR data take the form of long, thin tracts of very dense (13 m $\times$ 13 m) resolution measurements exhibiting considerable short scale structure, but often with large gaps in between. This finely spaced data over a relatively large spatial domain could allow a model to identify both long and short scale spatial variation if present in the observations, and both may be relevant when producing estimates—and their uncertainties—within the LiDAR observation transects as well as in the large gaps between them.

2.2. Prevalence of secondary education completion for women in Kenya

Sustainable Development Goal (SDG) 4 (United Nations, 2020) calls for improvements in secondary education to the point where everyone can complete their secondary education by 2030 regardless of their gender or the place where they live. Reliable spatial estimates of secondary education completion for young women are of particular importance to SDG 4. Yet in many developing countries, estimates of secondary education completion rely on complex, multistage household surveys (Li et al., 2019; Wagner et al., 2018) such as demographic health surveys (DHS) (USAID, 2019), multiple indicator cluster survey (MICS) (UNICEF - Statistics and Monitoring, 2012), AIDS indicator surveys (AIS) (DHS Program, 2019), and living standard measurement surveys (LSMS) (The World Bank, 2019).

Often, these household surveys are stratified by administration area and urbanicity; see for instance ICF International (2012). However, the classifications of urban or rural for the sampled clusters was made at the time of the last census, which at best takes place every 10 years, and the specific continuous spatial classifications of urbanicity used in the censuses are generally not made publicly available. This forces modelers to either ignore urbanicity or assume that the classification



(a) Forest canopy height (m) (b) Percent tree cover

Fig. 1. Forest canopy height in meters (a), and estimates of percent tree cover (b) in the Bonanza Creek Experimental Forest in Alaska, USA. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

remains accurate over large time spans, and to estimate urbanicity for unobserved locations based on proxy data such as population density (Paige et al., 2020; Wakefield et al., 2019). Because population averages over administrative areas are relevant for policymakers, any spatial model used in this context must be able to produce accurate averages in areas of varying size. In addition, socioeconomic indicators and other confounders may vary on multiple spatial scales. Models in this context will therefore need to accurately estimate correlations at spatial scales relevant for the administrative areas considered, for the potential confounders present, to be able to offset potential inaccuracies in urbanicity classification over short spatial scales near urban boundaries, and to reduce oversmoothing observed in Paige et al. (2020) when urbanicity is not accounted for. The response is population counts in this case, which are non-Gaussian, making LK poorly suited for this application.

It is in this context that we consider the prevalence of secondary education completion for young women in Kenya in 2014. Data are obtained from the 2014 Kenya DHS (KDHS, 2014) consisting of 1,612 clusters, each with official urban/rural designations, and age and educational achievement information for the sampled women within the cluster. The modeled response is the number of women aged 20-29 that have completed their secondary education. Since the response is non-Gaussian, LK cannot be applied, making this an ideal setting for ELK. We consider this application in more depth in Section 6. The dataset is illustrated in Fig. 2.

3. Methods

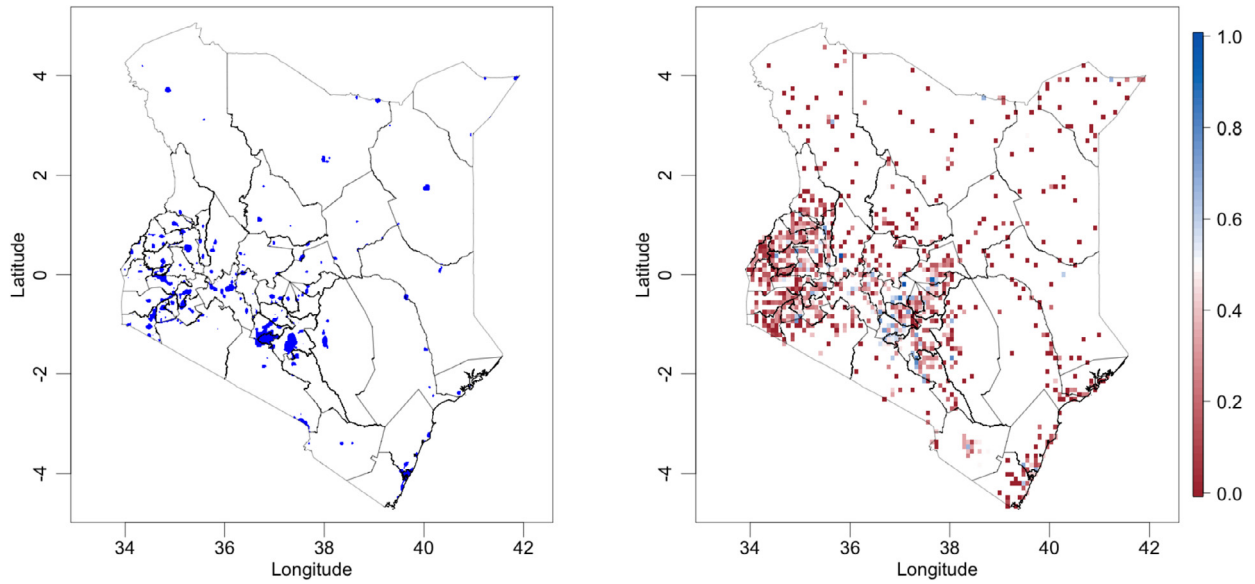
3.1. Background on LatticeKrig

LK, introduced by Nychka et al. (2015), has been used to analyze spatial datasets in a number of contexts (Bradley et al., 2013; Nychka et al., 2018; Thomas et al., 2014; Thomas, 2015). LK is a computationally efficient method for spatial modeling of the stochastic process $Y = \{y(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{D}\}$ in spatial domain $\mathcal{D}$ measured at observation locations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n \in \mathbb{R}^2$. The observation model is assumed to be Gaussian, with $y(\boldsymbol{x}_i)|\eta_i, \sigma^2_{\mathrm{N}} \sim \mathcal{N}(\eta_i, \sigma^2_{\mathrm{N}})$, $i = 1, \ldots, n$, where $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)$ are the linear predictors and σ^2_{N} is the nugget variance. The linear predictor is assumed to follow the linear model $\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{u}$, where $\mathbf{Z}$ is a $n \times p$ matrix where each column specifies a covariate, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is a vector containing the coefficients associated with the covariates, and $\boldsymbol{u} = (u(\boldsymbol{x}_1), u(\boldsymbol{x}_2), \ldots, u(\boldsymbol{x}_n))$ is the vector of values of the spatial Gaussian random field (GRF) u at the observation locations.

LK is characterized by the decomposition of u into a series of lattices of increasing spatial resolution, and over which increasingly fine basis functions are spaced:

$$u(\boldsymbol{x}) = \sum_{l=1}^{L} g_l(\boldsymbol{x}) = \sum_{l=1}^{L} \sum_{j=1}^{m(l)} c_j^l \phi_{l,j}(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{D} \subset \mathbb{R}^2.$$

Here L is a fixed, predetermined small number of lattice layers, usually between 2 and 4, and $g_1, ..., g_L$ are a series of smooth spatial functions associated with each lattice and composed of $m(1), ..., m(L)$ basis functions respectively. Each g_l is



(a) Urbanicity (b) Secondary education completion

Fig. 2. (a) Map of binary urbanicity classification in Kenya, where blue is urban and white is rural, and (b) 2014 empirical proportion of women aged 20-29 in Kenya that completed their secondary education.

respectively decomposed into a linear combination of basis functions $\phi_{l,1}, ..., \phi_{l,m(l)}$ with basis weights c_j^l, which are random variables.

Nychka et al. (2015) choose radial Wendland basis functions (Wendland, 1995), which have compact support. The basis functions are represented as $\phi_{l,j}(d) = \phi\left(\frac{d}{2.5\delta_l}\right)$, where $\phi(d) = (1-d)^6(35d^2+18d+3)/3$ for $0 \le d \le 1$, and 0 otherwise. Here δ_l is the layer l lattice cell width, and the factor of 2.5 ensures that the radius of each basis function is 2.5 times the respective layer lattice cell width. This overlap reduces artifacts in the predictive spatial means and standard errors (Nychka et al., 2015).

The basis coefficients for each layer respectively follow independent spatial autoregressive (SAR) models with mean zero multivariate normal distribution, $\mathbf{c}^l = (c_1^l, \ldots, c_{m(l)}^l) \sim \text{MVN}(\mathbf{0}, \alpha_l \sigma_S^2 \mathbf{B}_l^{-1} \mathbf{B}_l^{-T})$, where α_l determines the proportion of spatial variance σ_S^2 attributed to layer l. We also require $\sum_{l=1}^{L} \alpha_l = 1$. $\mathbf{B}_l$ is an autoregression matrix for layer l with elements $4+\kappa_l^2$ on the diagonal and up to four additional non-zero elements on each row corresponding to each neighbor, and with values of -1. As described in Lindgren et al. (2011), each layer l is a Gaussian Markov random field that approximates a Gaussian process with Matérn covariance function having smoothness $\nu = 1$ and effective spatial range approximately $\rho_l \equiv \sqrt{8}\delta_l/\kappa_l$. Hence, LK is similar to the SPDE model, except LK uses Wendland basis functions that are arranged in multiple lattice layers, and each layer is able to account for different spatial correlation scales. Also, Nychka et al. (2015) give theoretical results based on convolution processes (Higdon, 1998) showing that the LK class of covariance functions contains close approximations to Matérn covariances of arbitrary smoothness and range given sufficient lattice layers, and also give numerical results showing 3 or 4 layers is enough to closely approximate a variety of different covariance functions in practice. Note that Nychka et al. (2015) ensures the desired spatial variances at each point match the spatial variance σ_S^2 by numerically normalizing them. The interpretation of α_l as the proportion of variance attributed to layer l is not exact as the marginal variance of the different layers will vary depending on the values of κ_l.

Let $\mathbf{A}_l$ be the $n \times m(l)$ regression matrix from the basis coefficients for layer l to the basis function values at the coordinates of the observations so that $(\mathbf{A}_l)_{i,j} = c_j^l \phi_{l,j}(\boldsymbol{x}_i)$. We can then write the regression matrix from all basis coefficients to the values of the spatial effect at the observation locations as $\mathbf{A} = (\mathbf{A}_1 \ \ldots \ \mathbf{A}_L)$ so that $\boldsymbol{u} = \mathbf{A}\boldsymbol{c}$, where $\boldsymbol{c} = ((\boldsymbol{c}^1)^T \ \ldots \ (\boldsymbol{c}^L)^T)^T$. This means that the linear predictor can be written as $\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{c}$.

In the above formulation, LK requires p parameters for fixed effects, and $2L+1$ parameters for the covariance including the spatial variance σ_S^2, error variance σ_N^2, $L-1$ parameters for the layer weights, and L effective range parameters. It is sometimes assumed for simplicity that $\kappa_1 = \kappa_2 = \ldots = \kappa_L$, in which case the effective range of each layer is controlled exclusively by the layer resolution. Under this assumption, LK requires only $L+2$ covariance parameters.

3.2. A Bayesian extension to latent Gaussian models

We make two major additions to the formulation in the previous section: we allow for the model to be fit jointly with other structured random effects, and we allow for non-Gaussian responses. The model for the linear predictor is extended to

$$\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\beta} + \mathbf{A}\mathbf{c} + \sum_{i=1}^{K} \mathbf{M}_i \boldsymbol{\gamma}_i, \tag{1}$$

where each matrix $\mathbf{M}_i$ for $i = 1, \ldots, K$ is fixed, and defines a mapping to the observations from random effects collected in each vector $\boldsymbol{\gamma}_i$ associated with a component of the model such as temporal trends, space-time interactions, and other modeled effects. The vector $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_m^T)^T$ is assumed to follow a joint Gaussian distribution. Examples to help illustrate the importance of the $\sum_{i=1}^{K} \mathbf{M}_i \boldsymbol{\gamma}_i$ term are given in Section 3.3. Denote by $\boldsymbol{\theta}_\mathrm{M}$ and $\boldsymbol{\theta}_\mathrm{L}$ the vectors containing all model and family likelihood hyperparameters respectively. We can then formulate a latent Gaussian model in three stages. In stage 1, we have conditionally independent observations that may be non-Gaussian with likelihood $\pi(y(\boldsymbol{x}_i)|\eta_i, \boldsymbol{\theta}_\mathrm{L})$, $i = 1, 2, \ldots, n$. In stage 2, the latent model is a joint Gaussian distribution for $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{c})|\boldsymbol{\theta}_\mathrm{M}$. Lastly, in stage 3, we assign a prior $\pi(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_\mathrm{M}, \boldsymbol{\theta}_\mathrm{L})$.

The key computational contribution of this work is the adjustment of the LK formulation to be compatible with INLA, making Bayesian inference for the multiresolution latent Gaussian model computationally feasible. The combination is practically achieved by the development of the new model within the `INLA` package, which allows for the inclusion of the suite of models already included in INLA such as random walks which can be included temporally or for nonlinear models for covariate effects. We term the extended version of LatticeKrig, with computationally feasible inference, as extended LatticeKrig (ELK). Our approach exploits `GMRFLib`-library (Rue and Follestad, 2001) functions for sparse symmetric positive definite matrices based on methods described in Rue and Held (2005) when generating the latent coefficient precision and covariance matrices, and also precomputes relevant matrices and normalization factors whenever possible. Details on computations involved in our ELK implementation are given in Appendix S.1.

To ensure σ_S^2 can be approximately interpreted as the spatial variance and $(\alpha_1, \ldots, \alpha_L)$ as the proportion of spatial variance attributed to the layers, which is necessary to construct interpretable priors for these parameters, we normalize separately the SAR processes associated with each layer so that the variance of each g_l in the center of the spatial domain is $\alpha_l \cdot \sigma_S^2$. This requires the computation of normalization constants $\omega_1, ..., \omega_L$. Letting $\mathbf{A}_l^*$ be the $1 \times m(l)$ regression row vector that maps the layer l basis coefficients to the value of the basis functions at the center of the spatial domain, each ω_l can be calculated as: $\omega_l = (\mathbf{A}_l^* \mathbf{B}_l^{-1} \mathbf{B}_l^{-T} (\mathbf{A}_l^*)^T)^{-1}$. This is different from LK, since we only normalize the process to have variance σ_S^2 in the center of the domain rather than at every point. This has the advantage that it is faster computationally, and we find that if the lattice resolutions and buffers are chosen using the method discussed in the following paragraph, then the resulting process has spatial variance close to σ_S^2 across the whole spatial domain. In order to avoid matrix inversion and quadratic form computations each time $\mathbf{Q}$ is calculated, we precompute the mappings $f_l : \kappa_l \mapsto \omega_l$ using smoothing splines over a reasonable range of the values of κ_l.

3.3. Modeling additional structured random effects

To better understand the $\sum_{i=1}^{K} \mathbf{M}_i \boldsymbol{\gamma}_i$ term in (1), and to see why it adds so much generality to ELK, we could consider a number of simple examples below:

Nonlinear covariate effect: INLA natively supports random walk (RW) models that can be used for modeling nonlinear covariate effects. We might have n spatial observations, and a covariate of interest we wish to model using a RW of order one. We can divide the domain of the covariate into a set of 30 evenly spaced intervals, assigning a random walk coefficient to each interval. We then set $\boldsymbol{\gamma} \sim \mathrm{RW}(1)(\sigma_{\mathrm{RW}}^2)$ to be the assigned RW(1) coefficients, and let $\mathbf{M}$ be a $n \times 30$ matrix where $\mathbf{M}_{ij}$ is 1 if the covariate associated with observation i lies in the interval associated with RW coefficient j and 0 otherwise. We also include an intercept and linear trend in the covariate considered given by $\mathbf{Z}\boldsymbol{\beta}$. The linear predictor is then: $\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\beta} + \mathbf{A}\mathbf{c} + \mathbf{M}\boldsymbol{\gamma}$. This is essentially the same as the ELK model used later in Section 5.

Nonlinear covariate interactions: In addition to one dimensional RW models, INLA also supports 2D RW models. This allows not just for using random walks to model individual nonlinear covariate effects, but also for modeling nonlinear interactions between two covariates. Perhaps we have n binomial observations in space, and we wish to model them using two covariates that might have a complex nonlinear interaction. We could use a 2D random walk of order 1 in this case. Similar to the previous example, we may divide the domain of the covariates into a 30×30 grid, assigning a random walk coefficient to each grid cell. We set $\boldsymbol{\gamma} \sim \mathrm{RW2D}(\sigma_{\mathrm{RW2D}}^2)$ to be these 2D RW(1) coefficients, and let $\mathbf{M}_1$ be a $n \times 900$ matrix where $(\mathbf{M}_1)_{ij}$ is 1 if the covariates associated with observation i lie in the grid cell associated with RW coefficient j and 0 otherwise. Perhaps we believe there might be a spatial nugget effect (mean zero Gaussian variation) associated with each observation for non-Gaussian responses. Then we can define $\mathbf{M}_2 = \mathbf{I}_n$ and let $\boldsymbol{\gamma}_2 \sim \mathrm{MVN}(\mathbf{0}_n, \sigma_{\mathrm{nugget}}^2 \mathbf{I}_n)$. We also include an intercept and linear trends in the two covariates considered given by $\mathbf{Z}\boldsymbol{\beta}$. The resulting linear predictor is then: $\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\beta} + \mathbf{A}\mathbf{c} + \sum_{i=1}^{2} \mathbf{M}_i \boldsymbol{\gamma}_i$.

Repeated measurements through time: We might choose to model a set of T repeated observations at n spatial locations through time points $t = 1, \ldots, T$, making nT observations in total. If our covariates aside from β_0, the intercept, can be split into one set of covariates changing only in space and one set of covariates changing only in time, we could then model the fixed effects in space and time as $\mathbf{Z}_\mathrm{S}\boldsymbol{\beta}_\mathrm{S}$ and $\mathbf{Z}_\mathrm{T}\boldsymbol{\beta}_\mathrm{T}$ respectively for $n \times p_\mathrm{S}$ matrix $\mathbf{Z}_\mathrm{S}$ and $T \times p_\mathrm{T}$ matrix $\mathbf{Z}_\mathrm{T}$. Similarly, we

might assume that the spatial random effect varied only in space and the temporal random effect varied only in time. If the temporal trend is autoregressive order 1 (AR(1)), then we can set $\boldsymbol{\gamma} \sim$ AR(1) with correlation parameter ρ_{ar} and variance σ^2_{ar}, for T dimensional vector $\boldsymbol{\gamma}$. We could then define the linear predictor as,

$$\boldsymbol{\eta} = \mathbf{1}_{nT}\beta_0 + (\mathbf{1}_T \otimes \mathbf{Z}_S)\boldsymbol{\beta}_S + (\mathbf{Z}_T \otimes \mathbf{1}_n)\boldsymbol{\beta}_T + (\mathbf{1}_T \otimes \mathbf{A})\mathbf{c} + (\mathbf{I}_T \otimes \mathbf{1}_n)\boldsymbol{\gamma},$$

where '$\otimes$' represents the Kronecker product, and $\mathbf{I}_T$ is a $T \times T$ identity matrix so that $\mathbf{M} = \mathbf{I}_T \otimes \mathbf{1}_n$ adds the coefficients of $\boldsymbol{\gamma}$ identically to the coefficients of $\boldsymbol{\eta}$ associated with the corresponding time point. This model can be fit in the ELK framework with $K = 1$. Although not included in this model, interactions between the spatial and temporal effects could be included as well.

Note that this example shows how one might extend the ELK-T_u and ELK-T_U models introduced in Section 6 to space-time, since the urbanicity covariate used in that application is a spatial variable.

3.4. Recommended ELK settings

In LK, it is recommended that layer resolutions follow the relation $\delta_l = 2^{-(l-1)}\delta_1$, and when this relation is used in ELK under the assumption that $\kappa_1 = \ldots = \kappa_L$, we call this the 'fixed' model (ELK-F). Since the κ_l parameters are assumed to be equal to each other, ELK-F requires $L + 1$ hyperparameters, although more would be required if other latent effects were included in the $\mathbf{M}\boldsymbol{\gamma}$ term or for any likelihood family hyperparameters. A conservative guideline is for lattice resolutions to be at most a fifth of the effective range of the corresponding layer to avoid lattice artifacts and for accurate interpretation of the layer's effective range parameter. Since correlation lengths near the spatial domain diameter are very difficult to identify, we recommend choosing δ_1 to be finer than a fifth of the spatial domain diameter (typically around a twenty-fifth of the domain diameter), and setting $L = 3$ as a starting point. Spatial variation occurring on correlation scales smaller than δ_L will not be well-represented by $g(\cdot)$, so one must be careful when selecting basis layer resolutions so as to be computationally feasible, and also to be able to represent relevant correlation scales. The correlation lengths representable by ELK-F can differ by at most a factor of 2^{L-1}, which means that if the true spatial correlation function is a mixture of correlation functions with correlation lengths differing by factors larger than 2^{L-1}, ELK-F will not capture all relevant correlation scales in its predictions. One indicator of whether the correlation scales represented by an ELK-F model are adequate is to add an extra lattice layer, checking whether the estimated spatial covariances remains approximately the same. In particular, it is important to check that short scale correlations are not substantially changed by the added layer.

In the case where ELK-F model correlations change with the added layer, we propose to also consider a 'tailored' ELK model (ELK-T) with resolutions chosen for capturing variation at different spatial scales and with κ_l parameters allowed to vary for each layer. Since ELK-T allows the κ_l parameters to vary for each layer, it requires $2L$ hyperparameters, including $2L$ hyperparameters for the layer weights, total spatial variance, and correlation scales. Again, more would be required if other latent effects were included in the $\mathbf{M}\boldsymbol{\gamma}$ term or for any likelihood family hyperparameters. Just as for ELK-F, we recommend choosing δ_1 to be about a twenty fifth of the spatial domain diameter. However, since more hyperparameters are required, we recommend setting $L = 2$ and setting the fine scale resolution to be least as fine as a fifth of the minimum desired representable correlation scale that is computationally feasible. On most laptops, for about 1000 observations and for $L = 2$, a 100×100 lattice would be feasible for the fine layer, so we could choose δ_L to be about one ninetieth of the spatial domain diameter for ELK-T (adding a buffer of 5 basis elements in each direction makes a 100×100 lattice). Choosing δ_L to be larger would speed up computation at the cost of making very fine scale spatial correlations unable to be represented by the model.

For most practical purposes we see little reason to include more than 3 layers for ELK-T and 5 layers for ELK-F even if computation is feasible due to difficulty in model identification and lack of difference in predictive performance. However, more layers should be included, if feasible, for ELK-F in the case that relevant spatial scales differ by a factor larger than 2^{L-1}.

Fig. B.8 in Appendix S.2 in the supplementary material illustrates how the lattices might be arranged for a specific problem. In the figure and in Section 4 we use a buffer of 5 cell widths to avoid edge effects due to the zero boundary condition for the basis coefficients of each layer. The buffer size can be adjusted depending on the estimated effective correlation range for that layer.

3.5. Priors and specification

Expert knowledge on spatial scales at which dependence is expected could be used to choose appropriate lattice resolutions in ELK-T. Furthermore, the Bayesian formulation allows the inclusion of expert knowledge when setting priors for the interpretable parameters. For simplicity, we suggest a Dirichlet distribution of order L for the proportion of variances assigned to each layer, that is $\boldsymbol{\alpha} \sim \text{Dirichlet}(a_1, \ldots, a_L)$ with $a_l = 1.5/L$ for $l = 1, \ldots, L$ in order to place equal weight in the prior on each layer, and to ensure the prior is slightly concave for the sake of identifiability. Since the chosen prior concentration parameter is 1.5, the Dirichlet prior is only slightly more concave than the flat Dirichlet distribution that would result if the concentration parameter were 1. If the goal were to penalize differences from a Matérn covariance with smoothness ν, it would also be possible to apply Theorem 1 of Nychka et al. (2015) by setting $a_l \propto \delta_l^{2\nu}$, provided $\nu < 4 + 2K$

for Wendland basis functions of order K. On the spatial and nugget standard deviation we place penalized complexity (PC) priors satisfying $P(\sigma_S > 1) = 0.01$ and $P(\sigma_N > 1) = 0.01$, although this will depend on the context and prior information. See Simpson et al. (2017) for details on PC priors. As an alternative, it would be possible to select priors shrinking fine scale spatial variance relative to the large scale variance, or to place a joint PC prior on the layer variances using the method proposed in Fuglstad et al. (2020).

We propose setting independent inverse exponential priors for the effective range in each layer, where the effective range for layer l is computed as $\rho_l = \sqrt{8}\delta_l/\kappa_l$. Throughout this work, we set the median effective range of the coarsest layer at a fifth of the spatial domain diameter, determining any other effective range priors accordingly. However, the effective range priors can be customized to better suit the context as well as the expert knowledge of the modeler. For ELK-T, we recommend beginning by placing a prior on one layer's effective range, scaling priors for other layer effective range parameters proportionally to the lattice grid cell width δ_l for ELK-T. For ELK-F, a single κ parameter is estimated so that $\kappa = \kappa_1 = ... = \kappa_L$, and only one effective range parameter requires a prior. When placing priors on ELK-F or ELK-T effective ranges in this way, a prior on the effective range for one of the layers would therefore determine all effective range priors.

Note that if, under the ELK-T model, $\rho_1/\rho_L < 2^{k-1}$ for some integer $k > 0$, then the spatial correlation scales can be effectively represented by an ELK-F model with k layers. Hence, in addition to checking whether ELK-F spatial covariances are robust to adding an extra layer, the ELK-T model can be used as a test to check whether a given ELK-F model is able to represent all relevant spatial correlation scales.

A fully functional proof-of-concept implementation of ELK is freely available on Github at https://github.com/paigejo/ELK. Since it is implemented in `R` and not natively in `C++`, it does not reach its full potential in terms of speed.

4. Assessing performance under multiscale dependence

4.1. Simulation setting

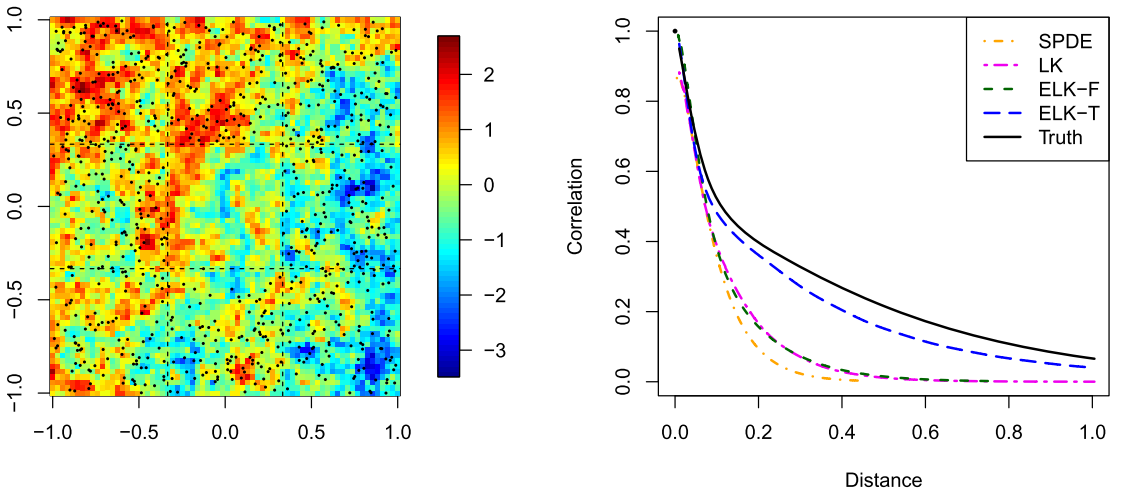
We first simulate a spatial GRF u on the square $[-1, 1]^2$. The GRF has the covariance function $C(d) = 0.5(C_1^*(d; 0.08) + C_1^*(d; 0.8))$, where $C_1^*(d; \rho) = (\sqrt{8}d/\rho)K_1(\sqrt{8}d/\rho)$ denotes the Matérn correlation (Stein, 1999) at distance d with smoothness $\nu = 1$ and effective spatial range ρ, and where K_1 is the modified Bessel function of the first order and second kind. The correlation function is plotted in Fig. 3 along with an example realization. The domain is then subdivided into a regular 3×3 grid, and we draw 800 observations at random locations, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{800}$, in the outer eight grid cells, but draw no observations within the central grid cell. We assume the unobserved latent process is $\eta_i = u(\boldsymbol{x}_i)$, and draw each observation $Y(\boldsymbol{x}_i)$ from $Y(\boldsymbol{x}_i)|\eta_i \sim \mathcal{N}(\eta_i, 0.1^2)$ for $i = 1, 2, \ldots, 800$. We fit several models, which we will describe in the next section, to the data, and generate predictions of the spatial process Y on a fine 70×70 grid. We also generate predictions of areal averages of the process Y for the nine subdivision areas, where areal averages are approximated numerically as averages of the values of Y on the 70×70 fine grid over each of the 9 areas. The whole procedure is repeated 100 times, and, for each realization, the predictions are scored in comparison to the truth. We choose to use Y as the process for comparing predictions rather than u so that comparison metrics are more similar to cross-validation, where only Y, and not u, is directly observed at the observation locations.

Computations for the simulation study are run on one of the Department of Mathematical Sciences' high performance computing servers at Norwegian University of Science and Technology (NTNU). The computing server has 47 24-core Intel(R) Xeon(R) Gold 6240R CPU 2.40 GHz central processing units (CPUs), and has a total of 376 Gb random access memory (RAM). We ran each model in serial using only 1 thread, although with only a single line of R code, more threads could have been used natively in INLA to speed up the SPDE and ELK calculations.

4.2. Prediction quality measures

The different spatial models are compared using three measures of predictive performance: root mean square error (RMSE), continuous rank probability score (CRPS) (Gneiting and Raftery, 2007), and empirical coverage of 80% prediction intervals. RMSE measures the average square error of the central predictions, whereas CRPS measures the integral of the squared difference between the empirical cumulative distribution function (CDF) and the predicted CDF, accounting for both error in the central predictions and uncertainty in those predictions. We also compared each model's runtime including setup, model fitting, and prediction and covariance parameter uncertainty calculations. We consider a number of different hold out schemes that are discussed in more detail in the sections that focus on the applications. For count responses, we rescale the counts to be empirical proportions with $y_i = y_i^*/N_i$ for observed count y_i^* with denominator N_i for observation $1 \leq i \leq n$. For each model considered, we calculate the measures as an average of its values over each held-out observation.

Unlike RMSE, CRPS is a strictly proper scoring rule (Gneiting and Raftery, 2007), and takes into account the accuracy of the central predictions as well as the calibration of the uncertainty. Smaller values are preferable. Prediction intervals at the 80% level are derived from the 0.1 and 0.9 quantiles of the predictive distribution. They are used to compute the prediction interval empirical coverage. For empirical proportions, and especially for small denominators, a fixed prediction interval will generally not provide the correct coverage even if the predictive distribution is correct due to the discreteness of the sample space (Geyer and Meeden, 2005). We therefore follow Geyer and Meeden (2005) by calculating *fuzzy* coverage instead, with



(a) One realization of the spatial field (b) True and estimated correlation functions

Fig. 3. (a) One of the 100 spatial field realizations. Black dots indicate the 800 observation locations and dashed lines indicate the 3×3 grid used for areal predictions. (b) True and estimated correlation functions averaged over 100 realizations.

details given in Appendix S.1. We have found that fuzzy coverage is much more precise, allowing us to be sure that observed over- or undercoverages are due to the accuracy of the predictive uncertainty rather than the discreteness of the CIs.

4.3. Models used in the simulation study

We use ELK-T with two layers: a grid of 14×14 basis knots and a grid of 126×126 knots over the spatial domain (not including the five knot buffer for each layer), which results in lattice resolutions of 0.154 and 0.016 respectively. In this case, the coarse and fine scale layers have at least five basis functions per 0.8 and 0.08 spatial units respectively. Further we use LK and ELK-F with three layers composed of 14×14, 37×37, and 53×53 lattice grids over the spatial domain with 0.154, 0.077, and 0.038 resolutions respectively. LK is fit using `LatticeKrig` in `R`, and for both LK and ELK-F, we use a single layer-independent parameter κ. Additionally, we fit an approximation to the Gaussian process with Matérn covariance and smoothness $\nu = 1$ using the SPDE approach with `INLA` (Lindgren et al., 2011; Lindgren and Rue, 2015). The mean triangular mesh segment length is approximately 0.0064 within the spatial domain.

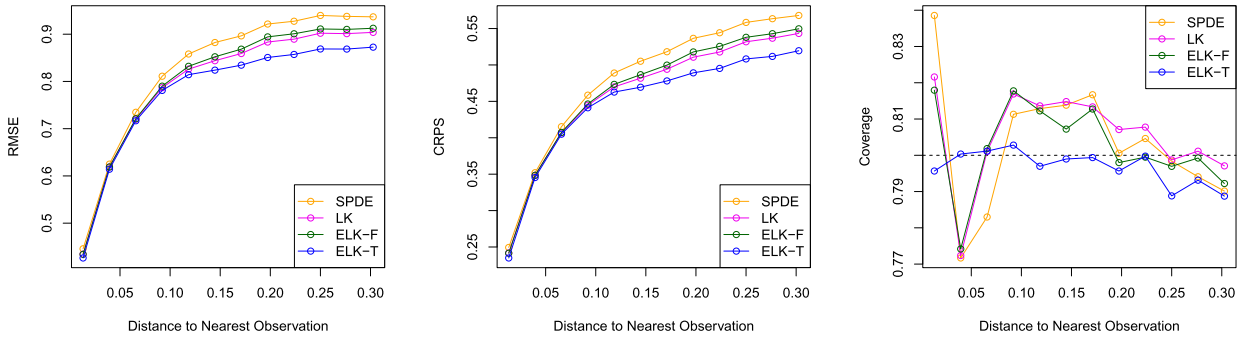
This gives a total of four models: ELK-T, ELK-F, LK, and SPDE. In all cases we use a PC prior for the nugget variance satisfying the tail probability $P(\sigma_\epsilon > 1) = 0.01$, and for the SPDE model we use the prior derived in Fuglstad et al. (2019) on the effective range and spatial variance. The median effective range for the prior is a fifth of the spatial domain diameter, and the spatial standard deviation again satisfies $P(\sigma_S > 1) = 0.01$. We use PC priors for the ELK-T and ELK-F spatial standard deviation also satisfying $P(\sigma_S > 1) = 0.01$, and use the effective range priors recommended in Section 3.2.

4.4. Results

For each realization and each of the Bayesian models, we generate 1,000 independent samples from the posterior distribution of Y (or conditional distribution in the case of LatticeKrig), estimate uncertainty in the parameters, and calculate covariance functions for 100 parameter samples out of the full set of 1,000. We used only 100 parameter samples when generating covariance function draws since for each draw the corresponding precision matrix for u must be inverted, which is especially computationally intensive for LatticeKrig since it does not take advantage of `GMRFLib` library functions for factoring sparse symmetric positive definite matrices, and since LK does not use ELK's simplified normalization scheme that precomputes normalization factors. In the case of LatticeKrig, we use the Hessian of the negative log likelihood to draw covariance parameter samples.

Fig. 3b) shows the central correlation function estimate for each of the models together with the true correlation function. The ELK-T model approximates the true correlation function over all distances well, while the other models strongly underestimate the spatial correlation after distance of 0.1, and have negligible correlation after distances of approximately 0.5.

Pointwise predictive scoring rules calculated by distance from prediction location to nearest observation are shown in Fig. 4. The RMSE and CRPS of ELK-T are the best in all of the distance bins. Differences in RMSE and CRPS among the models tend to increase as the distance to the nearest observation increases, but interestingly the differentiation is larger in the first bin than in the second bin. We believe this is due to the fact that ELK-T is able to capture the short range spatial correlation better than the three other models. The differences in RMSE and CRPS values for each model become increasingly large as



(a) RMSE vs distance to nearest observation

(b) CRPS vs distance to nearest observation

(c) 80% coverage vs distance to nearest observation

Fig. 4. Scoring rules calculated in bins depending on distance to nearest observation. The scores are averaged over 100 simulations, and include (a) RMSE, (b) CRPS, and (c) 80% uncertainty interval coverage.

Table 1
Scoring rules averaged over 100 simulated realizations and over a regular 70×70 grid of prediction locations across the entire spatial domain and areally integrated over all nine cells in the 3×3 regular grid partitioning the domain. Averages are calculated for each of the considered models. *Italics* indicate worse performance, **boldface** indicates better performance.

	RMSE	CRPS	80% Cvg	Runtime (min.)
Pointwise				
SPDE	*0.605*	*0.342*	**80**	**2.0**
LK	0.594	0.334	**80**	*51.1*
ELK-F	0.594	0.335	**80**	9.4
ELK-T	**0.587**	**0.329**	**80**	12.1
Areal				
SPDE	*0.137*	*0.056*	*75*	**2.0**
LK	0.121	0.051	77	*51.1*
ELK-F	0.125	0.052	77	9.4
ELK-T	**0.108**	**0.048**	**79**	12.1

distance to the nearest observation grows, indicating increasingly differing ability to accurately predict at locations far from the data.

Table 1 shows the summarized point and areal prediction scores. In terms of both pointwise and areal scores, the SPDE predictions have the worst RMSE, CRPS, and coverage in all cases, although the coverage of all the models matches the nominal level of 80% in the pointwise case. The coverage of the SPDE model is especially poor near the observations, indicating its inability to simultaneously capture short and long scale spatial correlations. Fig. 4 clearly demonstrates that even though SPDE achieves the correct nominal coverage overall, this is in spite of considerable over- and undercoverage depending on how far prediction locations are from the observations. There is also far more variability in coverage between bins for the SPDE model than ELK-T.

The SPDE model has the fastest runtimes. This is in part due to having an implementation built into and optimized in the `INLA` package, whereas ELK-T and ELK-F were implemented manually using the comparatively slow `rgeneric` framework (Gómez-Rubio, 2020, Ch. 11.3) intended for prototyping new models and special cases in `INLA`. However, the fact that the SPDE model requires only two hyperparameters excluding any family likelihood hyperparameters, compared to the four required in this case for ELK-F and ELK-T, further improves its computational performance. LK had the longest runtimes in large part due to the implementation of the predictive distribution sampling when calculating SEs. Drawing the 1,000 samples took over 33 minutes on average for LK, whereas drawing the same number of samples for the ELK-F model took under 2 minutes on average, and also included sampling over uncertainty in the hyperparameters.

The areal scores in Table 1 indicate a strong improvement from the SPDE model to ELK-F, and from ELK-F to ELK-T in terms of RMSE and CRPS. From the SPDE model to ELK-T, pointwise RMSE and CRPS scores improved respectively from 0.605 to 0.587 (3.0%) and from 0.342 to 0.329 (3.8%). However, in the integral prediction case, RMSE and CRPS scores improved respectively from 0.137 to 0.108 (21%) and from 0.056 to 0.048 (14%).

Table B.5 in Appendix S.2 in the supplemental material shows that the improvements in areal predictions are even larger when considering only the central grid cell, but Table B.6 in Appendix S.2 shows that there are improvements even when only the eight outer grid cells are considered. In summary, the results of this simulation study show that multiscale covariance models are essential both for accurate estimation of the covariance structure and for making pointwise and areal predictions when the true covariance function has both short and long scale structure.

5. Forest canopy height in Bonanza Creek experimental forest

5.1. Analysis

In this section, we consider the forest canopy height dataset introduced in Section 2.1. Due to exceptionally fine scale patterns in the observations, we use a similarly fine scale set of basis functions in the final lattice layer despite the larger computational costs. We consider 3 different models, including one LK model, an ELK-F model, and an ELK-T model. The LK and ELK-F models have an identical set of 4 layer lattices, where the coarsest lattice is 29×50 and the finest is 160×323, including the 5 cell buffer, resulting in 71,597 basis functions in total. The ELK-T model uses only the second and fourth layers from the ELK-F and LK models, where the second layer has dimension 48×89 including the 5 cell buffer, resulting in 55,952 basis functions in total for ELK-T. While this might seem like a relatively small reduction in the number of basis functions, the likelihood computations require $\mathcal{O}(m^{3/2})$ operations where m is the number of basis functions (Lindgren et al., 2011), so reductions in basis functions may result in nonlinear reductions in computational cost. All lattice layers are rotated 49.5° clockwise in order to better align with the spatial domain and thereby reduce the number of basis functions required for modeling.

The LK model assumes the response is Gaussian as in Finley et al. (2020b), with $y_i|\eta_i, \sigma_N^2 \sim \mathcal{N}(\eta_i, \sigma_N^2)$, whereas the ELK models assume the response follows a gamma distribution with mean $\exp\{\eta_i\}$ and variance σ_N^2 in order to avoid potentially problematic negative predictions, and due to a lack of heavy tailed behavior in the marginal distribution of the responses. All three models assume the following model for the linear predictor,

$$\eta_i = \beta_0 + u(\boldsymbol{x}_i) + f(\text{PTC}(\boldsymbol{x}_i)), \tag{2}$$

where β_0 is the intercept, u is the spatial field, and f is a function of $\text{PTC}(\boldsymbol{x}_i)$, the known percent tree cover at the spatial location of the i-th observation, $\boldsymbol{x}_i$. For LK, $f(\text{PTC}(\boldsymbol{x}_i))$ is linear with $f(\text{PTC}(\boldsymbol{x}_i)) = \beta_1 \text{PTC}(\boldsymbol{x}_i)$, whereas in the ELK-F and ELK-T models f is assumed to follow an order 1 random walk with 30 knots, $a_1, \ldots, a_{30}$, so that $f(a_j) - f(a_{j-1})|\sigma_{RW}^2 \sim \mathcal{N}(0, \sigma_{RW}^2)$, where the knots are spaced evenly throughout the range of percent tree cover in the spatial domain. Hence, in the ELK models f can be represented within the $\sum_{i=1} \mathbf{M}_i \boldsymbol{\gamma}_i$ term introduced in Section 3.2 as a $n \times 30$ matrix $\mathbf{M}_1$ multiplied by a vector of 30 random walk coefficients, $\boldsymbol{\gamma}_1$, where $(\mathbf{M}_1)_{ij}$ is 1 if a_j is the closest knot to $\text{PTC}(\boldsymbol{x}_i)$ and 0 otherwise. A sum to zero constraint is applied to the random walks to ensure identifiability.

For the ELK models, A PC prior is placed on the random walk increment standard deviation so that $P(\sigma_{RW} > 1) = 0.01$. The random walk is scaled to have unit variance in the null space of the covariance matrix to improve prior interpretability. We apply the recommended priors in Section 3.5 on the spatial range, variance, and layer weights.

All models are fit to a set of 105,504 observations from the dataset, lying along the majority of the main, longest data transect as well as on every other cross transect. The remaining 83,213 observations lie on every other cross transect as well as where they intersect with the main transect. Fig. C.11 in the supplement shows the locations of the in sample and out of sample data. The indices of the holdout indices can be obtained from the `BCEF` dataset in the `spNNGP` package.

Computations for this application are run on one of the Department of Mathematical Sciences' high performance computing servers at NTNU. This computing server has 27 14-core Intel(R) Xeon(R) CPU E5-2690 v4 2.60 GHz central processing units (CPUs), and has a total of 756 Gb RAM. It is different from the server used in the simulation study and other application due to the size of the dataset. We take advantage of the native support of INLA for multi-threaded computations by using a relatively modest 8 threads in the ELK-F and ELK-T models, whereas only 1 thread is used by LK due to its lack of native support for multithreaded computations. While we could easily use more threads when fitting the ELK models to speed up the computations further, 8 threads is a reasonable number for many laptops.

Central predictions and 80% uncertainty interval widths are plotted spatially in Fig. 5, and parameter estimates and uncertainties, where available, are given in Tables C.7-C.9 in Section S.3 in the supplement. Predictions and uncertainties are also given for ELK-F in the supplement in Fig. C.9. Some predictions and uncertainty interval widths lie outside of the [0 m, 40 m] range even though the maximum observed FCH was 37.16 m. For LK, all predictions outside of the range (about 1.5% of the predictions) are below 0, and are due to the Gaussian response assumption and basis function artifacts. For both ELK-F and ELK-T, a small proportion of the predictions and CI widths are above 40 m (less than 0.05% of them) due to basis function artifacts.

ELK-T gives central predictions that show longer correlation scales than LK, which signals improved ability to predict far from observations (this verified in the next Section via validation). The correlation scales influencing the predictions are especially noticeable in the uncertainties, which show stark contrast between uncertainties at locations within the in sample data transects and the locations immediately off of the data transects. While the effect of PTC is somewhat visible in LK's central predictions, they are much more noticeable in ELK's central predictions and ELK's uncertainties. This is due to both the nonlinear effect of PTC modeled by ELK, and ELK's log link, which causes the uncertainty to change as a result of the central prediction.

The sharp change in uncertainty interval widths over the LiDAR tracts versus between them is also due to the large proportion of spatial variance taken up by the finest resolution lattice layer, which is estimated by LK and ELK-T to be approximately 0.53 and 0.88 respectively. ELK-F, however, estimates the proportion of spatial variance in the finest layer to be only 0.34. This is due to the influence of the Dirichlet prior on the layer weights penalizing departures from the weights

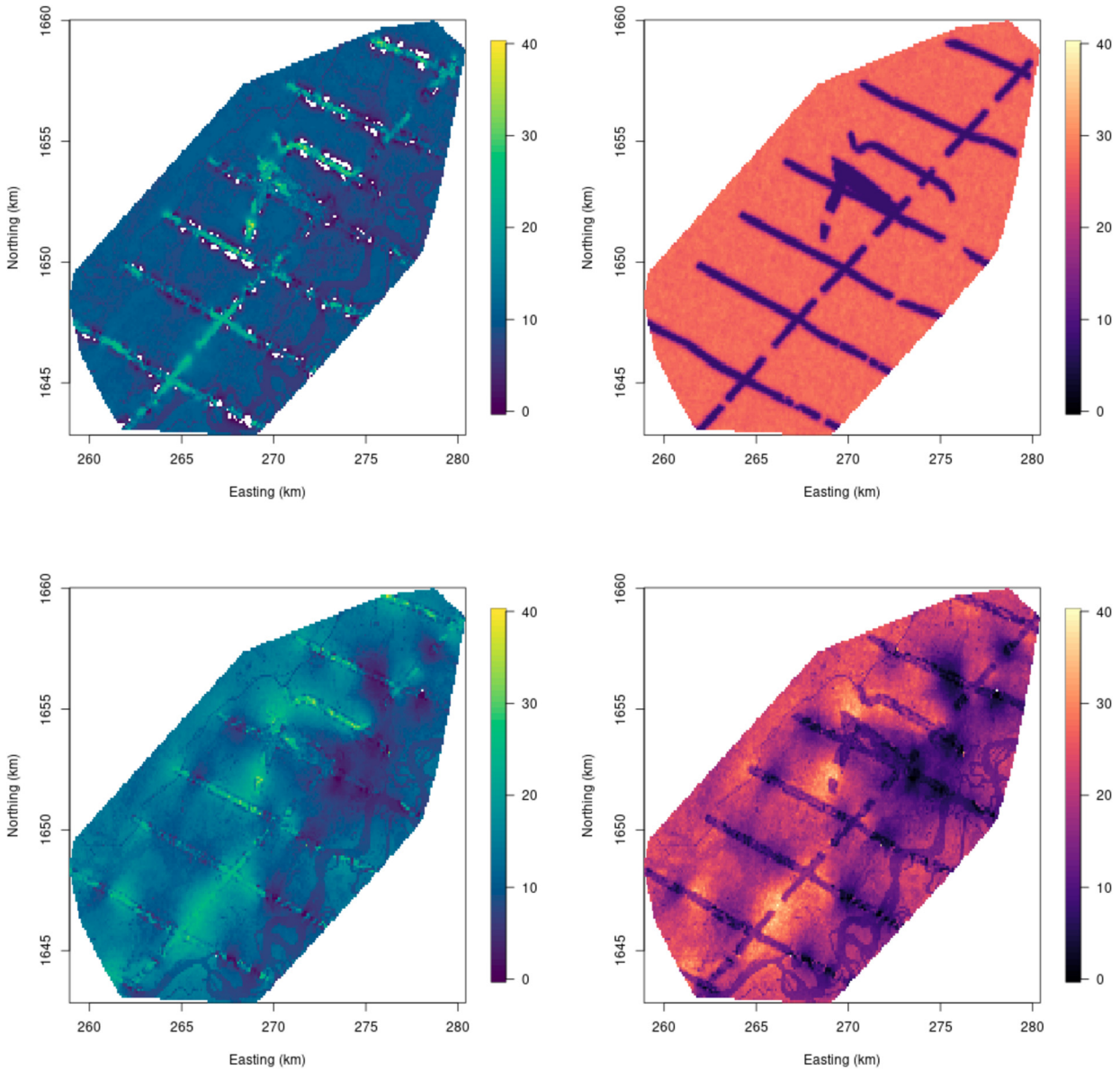
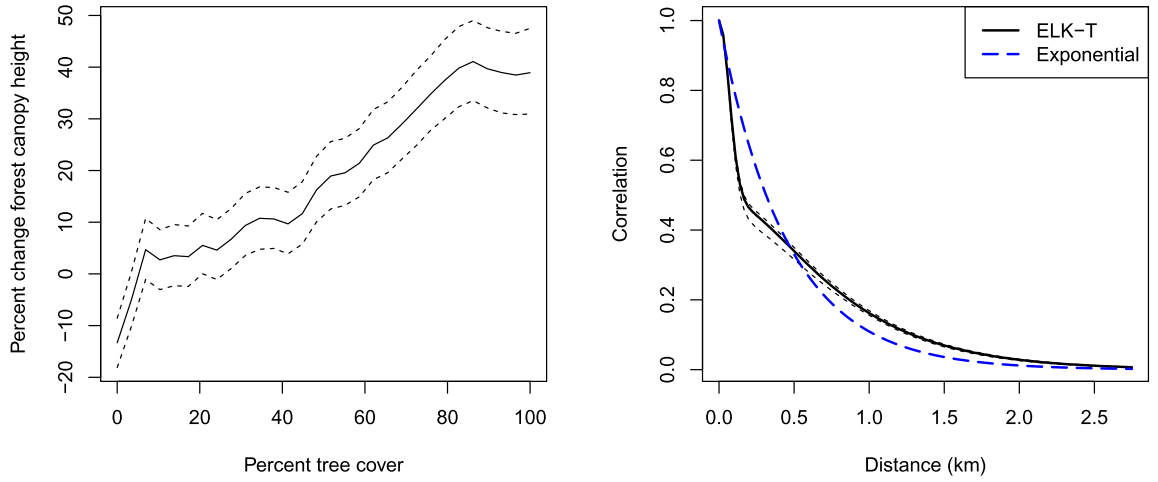


Fig. 5. Central predictions (left column) and 80% uncertainty interval widths (right column) for LK (top row) and ELK-T (bottom row).

all being equal. The effective correlation range of the finest lattice layer was estimated to be approximately 24 m for ELK-T. For LK and ELK-F, the effective correlation range of the finest lattice layer, obtained by dividing the estimated spatial range by 8, was estimated to be approximately 52 m and 47 m respectively.

We show the estimated correlogram according to ELK-T in Fig. 6 when compared against an exponential correlogram (i.e. a type of Matérn correlogram) with spatial range fit by minimizing the mean square error to the estimated ELK-T correlogram. Despite exponential covariances being used in a number of applications due to their ability to model short scale correlations (Berrocal et al., 2008; Liu et al., 2019), the ELK-T correlogram has higher correlation at both very short scales and very long scales, with less correlation near 0.2 km. This further suggests that short and long scale correlations may be important for predicting FCH, and helps to illustrate the practical importance of multiresolution spatial models and flexible covariance structures.

Fig. 6 shows the estimated effect of percent tree cover on forest canopy height according to the ELK-T model, and we found that the ELK-F estimated effect of percent tree cover was nearly identical. Overall, the variables appear to be positively associated, and while the association appears to be roughly linear on the latent modeled scale for most of the range of the percent tree cover variable, there is a sharp dropoff in the estimated effect for percent tree cover near 0%, and the estimated effect levels off above 80%, indicating nonlinearity.



(a) Effect of PTC (b) ELK-T correlogram estimate

Fig. 6. (a) Effect of percent tree cover with 80% pointwise credible bands, and (b) correlogram with 80% pointwise credible bands as estimated by ELK-T for the Bonanza Creek Experimental Forest dataset.

5.2. Validation

Validation scores and metrics for the three considered models are given in Table 2. A selection of the scores are plotted against distance to the nearest in sample observation in Fig. 7, where the boundaries of the distance bins over which scores are averaged are 0 m, 50 m, 100 m, 250 m, 500 m, 750 m,..., 2.75 km. Fig. C.10 in the supplement shows all of the considered scores plotted against distance to the nearest sample observation. The distances of the left out data to the nearest in sample observations are plotted spatially in the supplement in Fig. C.11, which shows that distances above approximately 1.75 km are clustered together in only 2 small areas on opposite ends of the same data transect. Hence, scores in the associated distance bins will be based on correlated residuals, making them imprecise.

We consider the bias, RMSE, CRPS, interval score (Gneiting and Raftery, 2007) for 80% uncertainty intervals (UIs), empirical 80% UI coverage, 80% UI width, computation time, and the number of threads used in the computation. Interval score penalizes both large UI widths and large distances between the response and the UI endpoint if the response lies outside for the UI. It is calculated via:

$$S_\alpha^{\text{int}}(l,u;y) = (u-l) + \frac{2}{\alpha}(l-y)\mathbf{1}\{y<l\} + \frac{2}{\alpha}(y-u)\mathbf{1}\{y>u\}$$

for a given significance level α, lower and upper UI thresholds given by l and u respectively, and response y. Like CRPS, the interval score is a strictly proper scoring rule, but interval score places greater stress on producing optimal UIs in terms of both their accuracy and width.

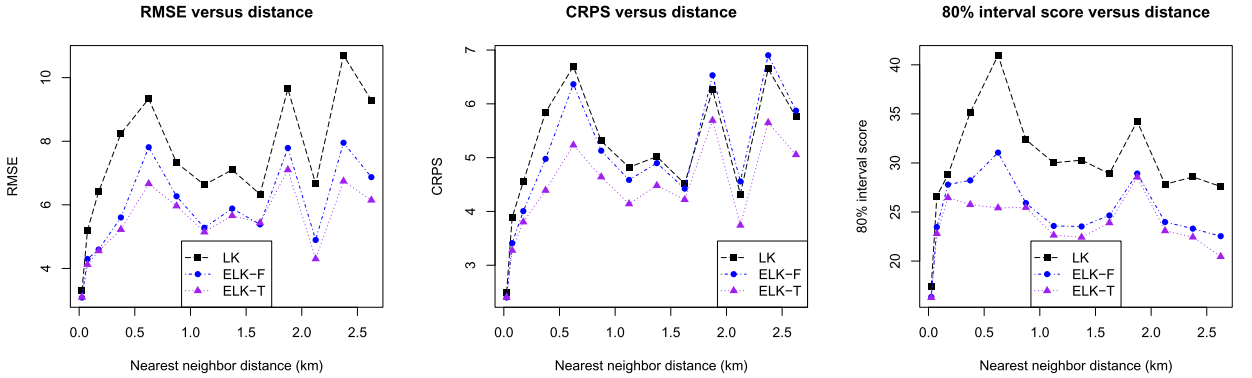
ELK-T and ELK-F performed the best and the second best respectively in every score and metric aside from 80% coverage, where LK performed the best with 84% coverage. The RMSE of ELK-F and ELK-T, 5.89 m and 5.52 m respectively, are comparable, and both significantly better than that of LK, which had a RMSE of 7.42 m. This shows the importance of ELK's ability to model the effect of PTC on FCH nonlinearly using a random walk model, where LK assumed PTC had a linear association with FCH due to its inability to include structured random effects (more specifically in this case, a random walk model of the first order). Interestingly, although LK had the best coverage, ELK-F and ELK-T had better interval scores due to their significantly lower UI widths of 22.1 m and 20.0 m in spite of their comparable coverages of 88% and 86% respectively, and where LK's UI width is 26.7 m.

Importantly, ELK-T outperforms LK in terms of RMSE, CRPS, and interval score for every considered distance bin, and ELK-F does as well with the sole exception of CRPS for distances beyond 1.75 km. It is important to note, however, that there are fewer left out observations in such high distance bins. This implies that ELK-T in particular is best able to flexibly account for multiple scales of spatial correlation in this example, both near and far from the observations. This is especially significant because of the many gaps between the data transects in which predictions must be made despite being potentially quite far from the observations.

ELK-T took only 12.4 hours to run compared to ELK-F's 24.7 hour runtime, and LK's 30.0 hour runtime. This shows the benefit of ELK-T's ability to remove unnecessary basis functions while providing greater flexibility with the basis functions retained, as well as the power of INLA's native support for multithreaded computations. Notably, ELK performs faster than LK despite ELK doing the extra work compared to LK of integrating over uncertainty in the covariance parameters. Although we have not recorded the runtimes, we found that, when run in serial, ELK-F was slower than LK, and ELK-T was only slightly faster than LK. This is most likely due to the optimized linear algebra operations in LK's likelihood calculations, which use

Table 2
Scoring rules and other metrics applied to left out data for the models fit to the Bonanza Creek Experimental Forest forest canopy height dataset. Units of bias, RMSE, CRPS, interval score, and 80% CI width are all in meters, computation time is in hours, and coverage is in percent. *Italics* indicate worse performance, **boldface** indicates better performance.

	Bias	RMSE	CRPS	80% Int. score	80% Cvg.	80% Width	Time	Threads
LK	*−5.93*	*7.42*	*5.30*	*32.5*	**84**	*26.7*	*30.0*	1
ELK-F	−3.20	5.89	4.97	26.0	*88*	22.1	24.7	8
ELK-T	**−2.79**	**5.52**	**4.42**	**24.0**	86	20.0	**12.4**	8



(a) RMSE vs distance to nearest observation

(b) CRPS vs distance to nearest observation

(c) 80% interval score vs distance to nearest observation

Fig. 7. RMSE (a), CRPS (b), and 80% interval score (c) versus the distance of the left out point to the nearest in sample observation for the three considered models.

a number of tricks to improve performance. Since the vast majority of ELK's linear algebra operations occur within INLA itself, this is unavoidable, but in this case, ELK-T's ability to use fewer basis functions in a robust way, and INLA's support for multithreaded computations led to ELK-T's comparatively fast computational performance.

6. Prevalence of secondary education completion

6.1. Analysis

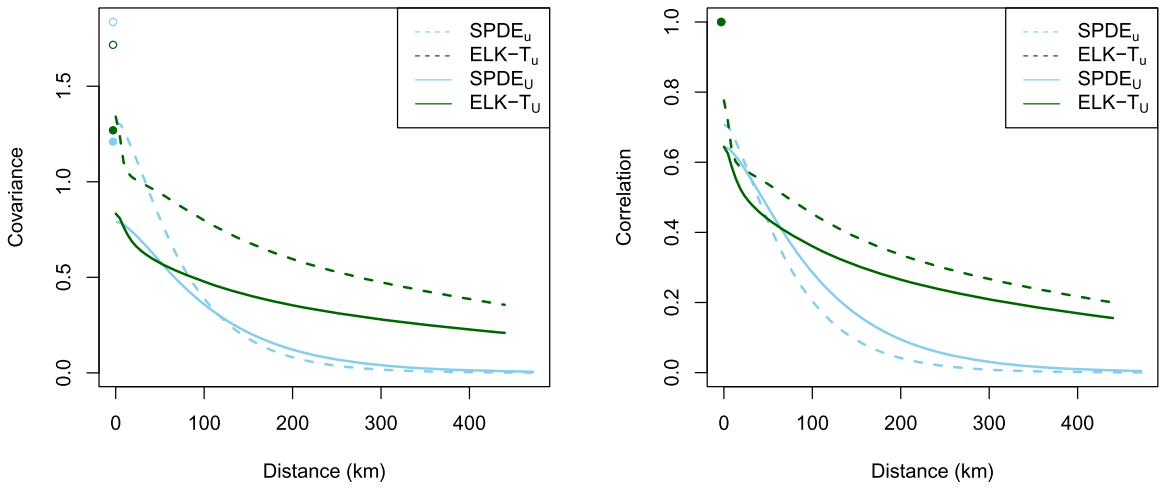
We return to the data introduced in Section 2.2: counts of secondary education completion for young women aged 20-29 in Kenya in 2014 using the 2014 Kenya DHS. The 2014 Kenya DHS household survey contains responses from individuals sampled from 1,612 clusters in 47 counties, each of which except Nairobi and Mombasa (which are both entirely urban) contain both urban and rural strata, making 92 strata in total. These 47 counties subdivide the 8 geographical provinces in Kenya. The response at cluster c, conditional on the probability of secondary education completion, $p(\boldsymbol{x}_c)$ at cluster spatial location $\boldsymbol{x}_c$, $c = 1, \ldots, 1612$, is modeled as, $Y(\boldsymbol{x}_c)|p(\boldsymbol{x}_c) \sim \text{Bin}(n_c, p(\boldsymbol{x}_c))$, where n_c is the total number of women aged 20-29 sampled in the cluster. The probability $p(\boldsymbol{x})$ is modeled on logit scale as,

$$\eta_c = \log\left(\frac{p(\boldsymbol{x}_c)}{1-p(\boldsymbol{x}_c)}\right) = \beta_0 + u(\boldsymbol{x}_c) + \beta^{\text{URB}}\mathbf{1}\{\boldsymbol{x}_c \in U\} + \epsilon_c, \quad c = 1, 2, \ldots, 1612, \tag{3}$$

with intercept β_0, spatial random effect $u(\boldsymbol{x}_c)$ with spatial variance σ_S^2, fixed effect for urban areas β^{URB}, and mean zero iid Gaussian cluster random effect ϵ_c with variance σ_ϵ^2. The indicator $\mathbf{1}\{\boldsymbol{x}_c \in U\}$ is 1 if $\boldsymbol{x}_c$ is in U, the set of urban areas in Kenya, and 0 otherwise. Equations for predicting at the point and areal level are given in Section S.4.1 in the supplement.

While it is possible to include other covariates, relevant demographic variables are not directly available for unobserved clusters, and pixel level estimates are modeled rather than observed directly. Moreover, the survey context of this problem makes inclusion of any pixel level covariates without potential for inducing bias in the areal estimates difficult. We therefore focus on urbanicity, which already requires precise aggregate estimates of urbanicity at the areal level in order to calibrate pixel level population density in urban and rural areas, and in order to generate the areal estimates with reasonable urban/rural weights. For more information on accounting for urbanicity in this context, see Paige et al. (2020).

The ability to model multiscale spatial correlations for this application is important in part due to the need to model short spatial correlations near urban areas (and changing urban boundaries), and long spatial correlations across rural areas. Additionally, it is important for policy makers to be able to make decisions at the administrative area level, making population averages over administrative areas necessary. However, LK is not applicable due to the binomial likelihood, so ELK will



(a) Covariogram Estimates (b) Correlogram Estimates

Fig. 8. (a) Spatial covariance, and (b) correlation estimates. The spatial nugget is plotted as the dots at zero distance with the color corresponding to the model given in the legends. Filled dots are plotted for models including urban effects, and unfilled dots are plotted for models without urban effects.

be necessary to model the multiscale correlations. Rather than compare with LK, we therefore compare two ELK models with two SPDE models. The four alternative models we consider are $\text{SPDE}_u/\text{SPDE}_U$, and $\text{ELK-T}_u/\text{ELK-T}_U$, where 'U' and 'u' respectively denote that urban effects are or are not included.

For ELK-T_u and ELK-T_U, the coarse lattice layer has 37 km resolution, while the fine layer resolution was set to be 5 km resolution in order to be able to capture sharp changes from urban localities to their rural surroundings. The SPDE model has an average triangular mesh segment length of approximately 15 km across the spatial domain. The spatial domain diameter is approximately $1,445$ km, so the prior median effective range was set to be one fifth of that, or 289 km, for the SPDE model and for the coarsest layer of the ELK models. Computations are run on the same computing server as in Section 4.

We use the Dirichlet prior for the layer weights recommended in Section 3.5, with concentration parameters set to $a_l = 1.5/L$ for $l = 1, \ldots, L$. We again place PC priors on the spatial and cluster variance parameters such that $\text{P}(\sigma_\epsilon > 1) = 0.01$ and $\text{P}(\sigma_S > 1) = 0.01$, except now the parameters should be interpreted on logit scale. All covariates except for the intercept are given noninformative Gaussian priors with zero mean and 0.001 precision, and the intercept is given an improper $\text{Unif}(-\infty, \infty)$ prior.

We test the sensitivity of our results to those where the Dirichlet concentration parameters are set to $a_l = 3/L$ and $a_l = 5/L$ for $l = 1, \ldots, L$ in the supplement in Section S.4.3. We find negligible sensitivity to the priors in the central predictions, and small, non-systemic differences in relative credible interval widths for the 2 ELK models, with average differences at the cluster level of only 3.2%. Part of these differences are due to Monte Carlo error from posterior sampling.

Central estimates for the correlation and covariance functions of the fitted models are shown in Fig. 8. Compared to the SPDE models, the ELK-T models incorporate more long scale spatial correlation while also modeling short scale correlations with more subtlety as shown by their long tailed covariance and correlation functions with sharp downward trends at small spatial distances. Including urbanicity as a covariate substantially reduces the spatial variance for all models, and also reduces the variance of the spatial nugget. We find that including an urban effect explains spatial variation at both short and long scales, because sharp changes due to urban/rural boundaries are accounted for, as well as long scale correlations across rural regions. We see this effect in the estimated correlation function of the ELK-T models, where the magnitude of the relatively sharp downward trend in correlation at small distances decreases when the urban effect is included, and where the long tail shortens slightly as well. Since the likelihood under Matérn correlation (or Matérn approximations like the SPDE model) is primarily affected by short correlation scales, the sharp changes in education due to changes in urbanicity rather than the long scale correlations induced by large areas being rural drive the correlation function estimate. Hence, including the urban effect in the SPDE model removes some of the otherwise unmodeled spatial correlation at short spatial scales, increasing the estimated effective range. It is worth noting, however, that even with an urban effect, the ELK-T_U model covariance estimates are still different from those in the SPDE_U model at both short and long scales.

In Fig. 9 we give pixel level predictions at the 5 km $\times$ 5 km resolution of secondary education prevalence as well as relative credible widths, which we define as credible widths divided by the corresponding central estimates. Areal predictions are created based on aggregation of pixel estimates weighted by population density as in (5-6) of Paige et al. (2020), except we leave out cluster effects by setting them to 0 rather than integrating over them as done in (7) of Paige et al. (2020). As in Paige et al. (2020), we use a 1 km $\times$ 1 km population density grid over Kenya interpolated assuming constant growth rate using WorldPop data (Stevens et al., 2015; Tatem, 2017) from 2010 and 2015 to get a 2014 estimate. The resulting adjusted population density surface is available on GitHub at https://github.com/paigejo/U5MR/blob/master/popGridAdjustedWomen.

Table 3
Scoring rules calculated for each model using leave one province out and stratified 8-fold cross validation. Scores are averaged for each province, over urban areas, and over rural areas. *Italics* indicate worse performance, **boldface** indicates better performance.

	RMSE	CRPS	80% Cvg	Width
Leave One Province Out				
$SPDE_u$	*0.238*	*0.129*	76	0.52
$SPDE_U$	0.224	0.119	*74*	**0.47**
ELK-T_u	0.234	0.125	**77**	*0.53*
ELK-T_U	**0.223**	**0.117**	**77**	0.49
Stratified 8-Fold				
$SPDE_u$	*0.226*	*0.119*	73	0.46
$SPDE_U$	**0.218**	0.114	*72*	**0.42**
ELK-T_u	0.223	0.117	**77**	*0.49*
ELK-T_U	**0.218**	**0.113**	75	0.45

RData. Locations are determined as being urban or rural based on a population density threshold set for each county to ensure the correct proportion of the population is urban/rural using the same method as Paige et al. (2020). Predictions and relative credible widths aggregated to the county and province levels using the population density surface are shown in Appendix S.4.2 in Figs. D.12 and D.13. Tables of the county level and province predictions for the models with urban effects as well summary statistics for the model parameters are given in Appendix S.4.2. in Tables D.11-D.13.

The pixel level predictions show nearly indistinguishable differences in predictions and uncertainties between the $SPDE_U$ and ELK-T_U models, but much more significant differences in the predictions between the $SPDE_u$ and ELK-T_u models. In particular, the ELK-T_u model shows reduced spatial oversmoothing near urban areas, and higher uncertainties overall. These uncertainties reflect that an important confounder in urbanicity is not included as a covariate. The reduction in oversmoothing is especially noticeable in the north and east counties with large rural areas and spatially concentrated urban areas, although there are reductions in oversmoothing in other areas as well. The differences between the models without urban effects, and the similarities between the models with urban effects are further highlighted in the pair plots in Fig. 10, which shows the predictions of the $SPDE_u$, ELK-T_u, and $SPDE_U$ models sequentially move towards the predictions of the ELK-T_U.

That the $SPDE_U$ and ELK-T_U predictions are essentially indistinguishable lends credence to ELK's predictions by showing they agree with the SPDE model, an established model when fit via INLA, indicating that MCMC can be effectively avoided. It also suggests that there is little identifiable spatial covariance at very short scales that is not already accounted for by urbanicity, and that the aggregate effect of remaining spatial confounders probably varies smoothly over medium to long spatial scales, or scales shorter than 5 km. More broadly, this suggests that ELK could be used as a tool to help identify important spatial confounders if they vary on the same spatial scales as ELK's primary estimated spatial correlation scales.

6.2. Validation

We use two different schemes to validate our models: leave one province out, and stratified, eight-fold cross validation (CV). In the leave one province out scheme, we calculate scoring rules based on the predicted distributions of the left out clusters in each of the 8 provinces consecutively, averaging the scores within each province, and then averaging the province scores to get the final reported scores. In the stratified, eight-fold CV, we randomly partition the set of clusters in each of the 92 strata (47 counties except Nairobi and Mombasa have both urban and rural, while Nairobi and Mombasa are entirely urban) into eight roughly equal sized folds. The former approach has the advantage of being able to measure predictive performance farther from nearby in sample observations, while the latter measures predictive performance closer to nearby in sample observations. We make sure that, for a given stratum, the difference between the number of clusters in each fold is different by at most one, and that which folds get more clusters than others is random. We choose eight folds since the smallest stratum has only eight clusters. The two different validation schemes give an idea of both short and long scale predictive errors due to the distribution of how far away left out clusters are from in sample observations. The leave one province out scheme better identifies long scale errors, and the stratified CV better identifies short and medium scale errors. The boundaries of the 8 provinces are plotted in Fig. 9 along with county boundaries.

The results from the leave one province out and the stratified CV are given in Table 3. The ability of the ELK-T model to account for more flexible spatial covariance structures than the SPDE model leads to as good or better predictions from RMSE, CRPS, and coverage standpoints, although the improvement is clearly greater when the urban effect is absent in the model. Improvements were especially noticeable in the leave one province out CV, where long range correlations mattered more, and relative improvements were greater for CRPS than for RMSE. For leave one province out CV, RMSE improved by 1.7% when urban effects were not included in the SPDE and ELK-T models respectively, and by 0.4% otherwise, while CRPS improved by 3.1% when urban effects were not included, and by 1.7% otherwise. The $SPDE_U$ model had the worst coverage with 74%, and both ELK-T models tied for the best coverage with 77%.

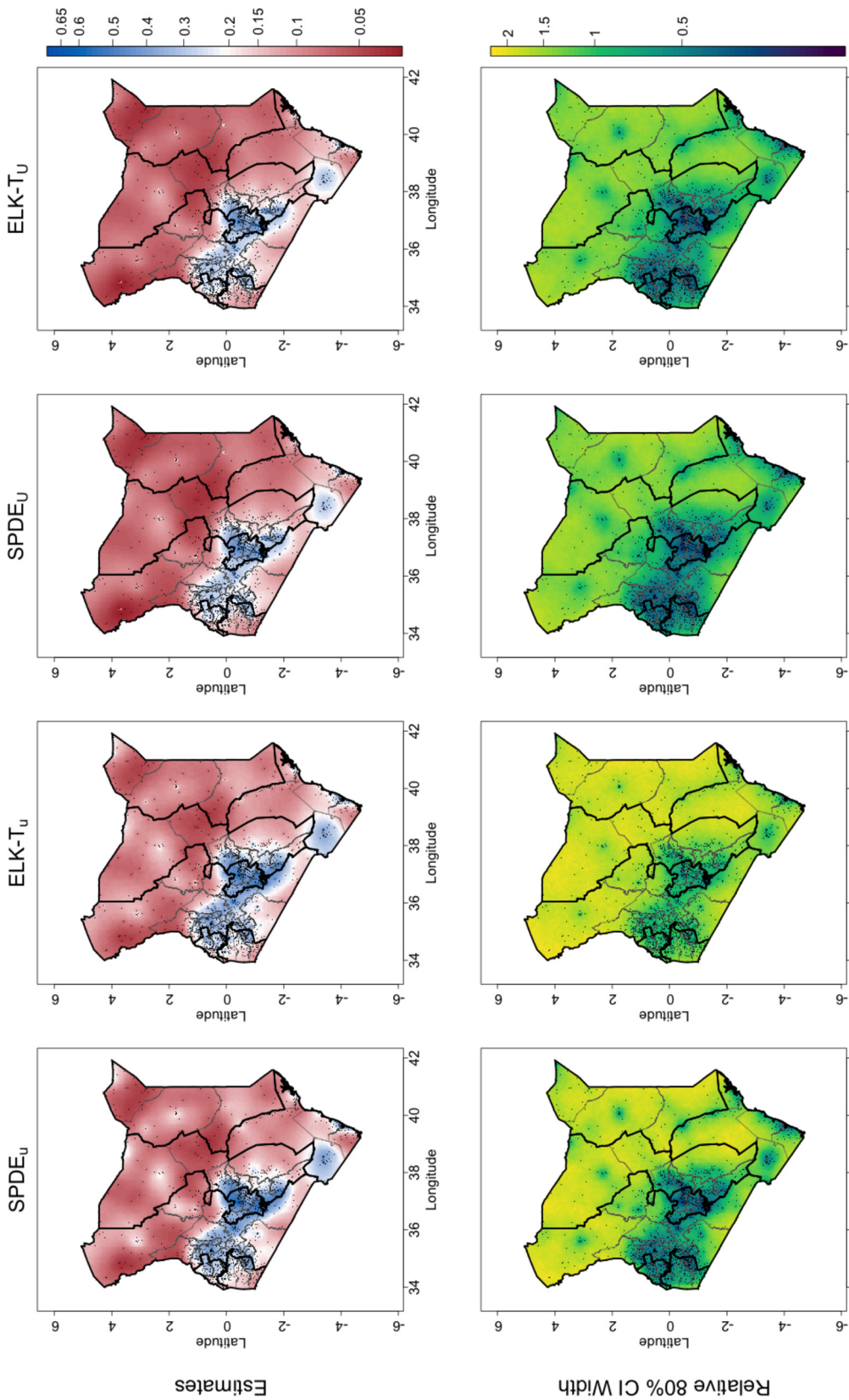


Fig. 9. Central 5 km × 5 km pixel level predictions (top row) and relative 80% credible interval widths (bottom row) of secondary education prevalence for young women in Kenya in 2014. Models with subscript 'U' and 'u' respectively do and do not include urban effects. Observation locations are plotted as black dots, provinces as thick black lines, and counties as thin gray lines.

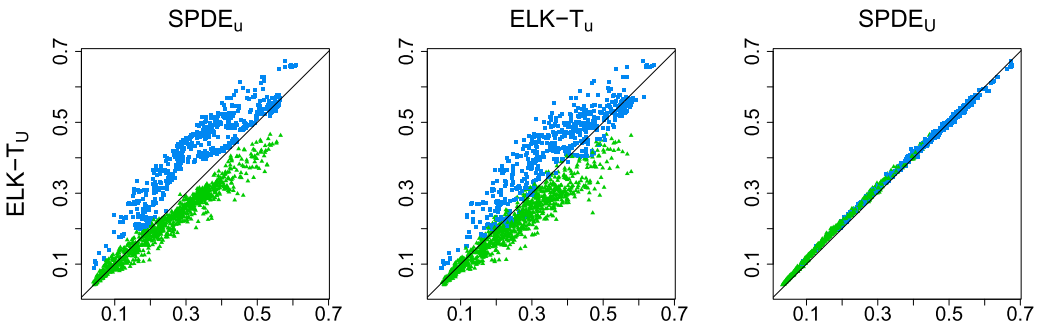


Fig. 10. Pair plot of the cluster level estimates comparing the considered models' estimates of secondary education prevalence to the ELK-T_U. The '▲' symbols are rural clusters, while '■' symbols are urban clusters.

7. Discussion

The LK model introduced by Nychka et al. (2015) attempts to answer the question of how to flexibly model spatial covariance at different spatial scales in a way that is computationally feasible even for large datasets. LK also provides a way to interpret its modeled spatial process that may be valuable in practical contexts: as a linear combination of independent spatial processes, each with its own spatial range. However, LK has a number of limitations that we address in this work. For instance, LK is unable to model non-Gaussian observations, which severely limits the datasets it can be applied to. LK also is unable to integrate over uncertainty in its covariance parameters (in fact, calculating the uncertainty in the LK parameters and covariance estimate is surprisingly difficult in the `LatticeKrig` R package), and does not support penalizing covariance parameter values departing from a simple base model, which may limit the robustness of its predictive uncertainties considering the flexibility of the covariance model. LK does not natively support multithreaded parallel likelihood calculations, limiting its ability to quickly analyze large datasets. LK also does not support the inclusion of structured and unstructured random effects in the model, which limits its ability to model spatiotemporal data using random effects for temporal terms, random effects included at the areal level, and nonlinear covariate effects, among other examples.

Our proposed ELK model extends LK, and is able to model a much more general set of responses including exponential family responses, and even some non-exponential family responses such as the betabinomial distribution. Additionally, it is implemented in INLA, which avoids the computational cost of MCMC, provides native multithreading and support for parallel computation, allows for easy integration over covariance parameters and covariance parameter uncertainty estimation, and enables multiresolution spatial modeling within a more general class of models: latent Gaussian models. Moreover, we find that, computationally, ELK can perform faster than LK when uncertainty in the predictions and covariance parameters is desired, even at times when not taking advantage of INLA's ability to perform multithreaded computations. This is in part due to ELK's ability to tailor its lattice resolutions to the context, removing unnecessary lattice layers, and allowing for the remaining layers to have independent spatial ranges without necessarily increasing the number of parameters. Notably, ELK has access to the full suite of models that can be fit in INLA such as nonlinear random effects models for time series or covariates.

We show using a simulation study that the ability of LK and ELK to model spatial covariance flexibly can be important for predictive performance at both short and long scales. We find that, while short scale dependence is most important for point level predictions near observations, long scale dependence can matter more when making predictions in data sparse regions, and when making areal predictions.

We apply ELK to a 188,717 observation forest canopy height dataset exhibiting both short and long scale structure in the data, and show that, like LK, ELK can be applied to relatively large datasets. We find that ELK's flexibility in tailoring the lattice layers to the application can help reduce the number of basis functions as well as computation time, while also allowing for greater flexibility in the ability to model multiscale spatial structure. ELK is also able to jointly account for nonlinear covariate effects modeled via random walks. Sharp changes in the uncertainties at the edge of the LiDAR tracts, contrasted by gradual changes in uncertainties in between the tracts showed how LK and ELK were able to simultaneously model the short and long scale spatial structure in the dataset.

We also apply ELK to a 2014 Kenya DHS dataset with information on the prevalence of secondary education for women aged 20-29 in 2014, a setting where multiscale spatial correlation is of particular importance due to the tendency for the response to change sharply near urban boundaries and vary little over large rural areas, and where spatially aggregated predictions for administrative areas—both large and small—are necessary. Under a binomial likelihood, however, LK is not applicable. We find noticeable reductions in spatial oversmoothing relative to a SPDE model, especially when confounding by urbanicity was accounted for by a fixed effect in the linear predictor. Evidence of short scale spatial confounding was present in the estimate of the spatial correlation function in the ELK model with no urban effect, indicating that ELK can make predictions more robust to spatial confounding as well as be indicative of the spatial scales at which spatial confounding is occurring. This in turn can suggest what variables should be included as covariates, and as an informal check for spatial confounding. In general, it is very difficult to tell whether an unmeasured covariate is confounding results,

but ELK provides at least a modicum of insurance against this. Since DHS household surveys tend to consist of clusters that are spatially concentrated in urban areas and sparsely distributed in rural areas, this is an application that ELK is well suited for.

In a spatial context where identifiability is already difficult, spatial confounders and the flexibility of LK when layer correlation ranges are allowed to independently vary further reduces identifiability. While the `LatticeKrig` package does not support model penalization, ELK supports penalization in the form of the chosen priors. In particular, the Dirichlet prior on the lattice layer variances and the negative exponential priors on the spatial effective range parameters can be chosen to penalize departure from an approximate Matérn covariance using theoretical relationships established in Nychka et al. (2015), or departure from equal layer variances. ELK is also able to integrate over parameter and hyperparameter uncertainty, which is especially important for modeling flexible and uncertain covariance structures. By allowing for the incorporation of priors and integration over parameter uncertainty without significant reductions in computational performance, ELK's Bayesian framework is a valuable extension over standard LK.

A potential future extension to ELK would be to incorporate joint PC priors (Fuglstad et al., 2020) on the lattice layer weights and variances in order to penalize model complexity in a more principled way rather than using a Dirichlet prior, although this would come at greater computational cost.

Depending on the context, one may choose to select lattice resolutions that are independent of each other rather than changing by a factor of two from one layer to the next as in standard LK. In both the illustrative example and the applications, we found that forcing each consecutive layer to have double the resolution along each dimension made modeling the fine and long scale correlations simultaneously difficult from a computational perspective. This is due to the number of hyperparameters and basis functions required to model spatial correlations on scales differing by factors of 10 or more. In such situations, we advocate for tailoring the resolutions of each lattice to enable them to model a set of effective ranges of interest.

We find that it may be advantageous to choose ELK over a Matérn model in cases where there is reason to believe spatial correlations are operating on multiple spatial scales, possibly due to spatial confounding. It may also be worthwhile to choose ELK or LK when spatial correlations are non-Matérn, when making areal predictions for areas of many different shapes and sizes, and when making predictions using observations with large spatial gaps such as in LiDAR or satellite data.

Acknowledgements

JP was supported by The National Science Foundation Graduate Research Fellowship Program under award DGE-1256082. GW and AR were supported by project number 240873 from the Research Council of Norway. JW was supported by the National Institutes of Health under award R01 AI029168-30.

Appendix A. Relevant correlation scales for spatial integration

Long-range correlations are especially important when calculating predictions of certain areal averages. With a 'back of the envelope' calculation one can calculate the variance of a predicted spatial integral over a disk with radius R. Let $\hat{r}(d)$ be the estimated covariance, and let $r(d)$ be the true covariance such that,

$$\hat{r}(d) = r(d) + e(d),$$

so e is the error in the covariance estimate a function of distance. Then if we denote the disk by A, and the true spatial field with $g(\mathbf{x})$, the variance of our spatial integral under the predictive distribution is:

$$\begin{aligned}\widehat{\mathrm{Var}}(g(A)) &= \int_A \int_A \widehat{\mathrm{Cov}}(\mathbf{u}, \mathbf{v})\, d\mathbf{u}\, d\mathbf{v} \\ &= \int_A \int_A r(||\mathbf{u} - \mathbf{v}||) + e(||\mathbf{u} - \mathbf{v}||)\, d\mathbf{u}\, d\mathbf{v} \\ &= \mathrm{Var}(g(A)) + \int_A \int_A e(||\mathbf{u} - \mathbf{v}||)\, d\mathbf{u}\, d\mathbf{v}.\end{aligned}$$

Let D be the random distance between any two points chosen in the disk with independent uniform distributions. Then Tuckwell (2018) shows the density of D is:

$$p_D(d) = \begin{cases} \frac{4d}{\pi R^2}\left(\arccos\left(\frac{d}{2R}\right) - \frac{d}{2R}\sqrt{1 - \left(\frac{d}{2R}\right)^2}\right), & 0 \le d \le 2R \\ 0, & \text{otherwise.}\end{cases}$$

Hence,

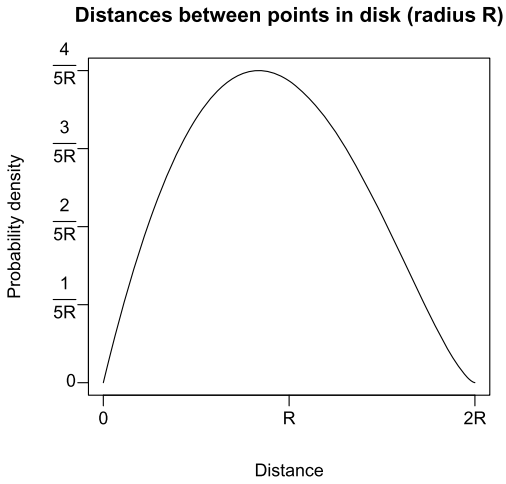


Fig. 11. The distribution of distances between points uniformly distributed on a disk of radius R.

$$\widehat{\text{Var}}(g(A)) = \text{Var}(g(A)) + \int_0^{2R} e(D) \cdot \frac{4d}{\pi R^2} \left(\arccos\left(\frac{d}{2R}\right) - \frac{d}{2R}\sqrt{1-\left(\frac{d}{2R}\right)^2} \right) dD.$$

Fig. 11 shows that the density $p_D(d)$ roughly parabolic with peak just under R (approximately $0.834R$), and has zeros at 0 and $2R$. Because of this, errors in very short and very long-range correlations are less relevant than errors in the assumed correlation function at the spatial scale near the radius of the area over which we integrate, R, when calculating predictive uncertainties. This is of course not the full story, since the covariance structure conditional on the data will not be so neatly stationary and isotropic, and will likely have shorter spatial range. At the same time, we believe this shows greater emphasis must be placed on long range spatial correlations when producing area level predictions, especially in large areas.

Appendix B. ELK sparse matrix computations

The computational performance of our implementation of ELK within *inla* is almost entirely determined by how quickly the sparse precision matrix of the basic coefficients $\mathbf{c}$ can be generated. As such, we precompute any information for this task that will improve the performance. Recall that the basis coefficients for each layer follow independent SAR models with mean zero Gaussian distribution, $\mathbf{c}_l \sim \text{MVN}(\mathbf{0}, \alpha_l \sigma_S^2 \mathbf{B}_l^{-1} \mathbf{B}_l^{-T})$, with,

$$\mathbf{B}_{l,i,j} = \begin{cases} 4+\kappa_l^2, & i=j \\ -1, & i \in N_l(j) \\ 0, & \text{otherwise,} \end{cases}$$

where $N_l(j)$ is this set of indices of lattice knots in layer l neighboring lattice knot i. The precision matrix for layer l, $\mathbf{Q}_l$, can therefore be represented as,

$$\mathbf{Q}_l = \frac{\omega_l}{\alpha_l \sigma_S^2} \left(\kappa_l^4 \mathbf{I}_{m(l)} - \kappa_l^2 (\mathbf{D}^l + (\mathbf{D}^l)^T) + (\mathbf{D}^l)^T \mathbf{D}^l \right),$$

for matrices,

$$\begin{aligned} \mathbf{D}^l &= \mathbf{D}_x^l + \mathbf{D}_y^l \\ \mathbf{D}_x^l &= \mathbf{I}_{m_y(l)} \otimes \nabla^2_{m_x(l)} \\ \mathbf{D}_y^l &= \mathbf{I}_{m_x(l)} \otimes \nabla^2_{m_y(l)}, \end{aligned}$$

where $m_x(l)$ and $m_y(l)$ are the number of basis functions in the horizontal and vertical directions of layer l, $\mathbf{I}_{m_x(l)}$ and $\mathbf{I}_{m_y(l)}$ are $m_x(l) \times m_x(l)$ and $m_y(l) \times m_y(l)$ identity matrices respectively, and '$\otimes$' denotes the Kronecker product. Note that the variance normalization factor ω_l is a function of κ_l, although we leave out this dependence in the notation for simplicity. We can therefore precompute $\mathbf{D}^l + (\mathbf{D}^l)^T$ and $\mathbf{D}^l)^T \mathbf{D}^l$ in order to calculate $\mathbf{Q}_l$ as quickly as possible for each chosen value of κ_l.

Since there is no exact closed form solution for the functions $f_l : \kappa_l \mapsto \omega_l,\ l = 1, \ldots, L$, they are approximated using monotonic smoothing splines (Hyman, 1983) fit on a log-log scale over a set of reasonable effective ranges for each layer.

Throughout this paper, the effective ranges used for fitting f_1 vary from a fifth of the first layer lattice width to the diameter of the spatial domain, and the effective ranges used when fitting subsequent f_l shrink proportionally with the corresponding lattice widths. Hence, if w is the diameter of the spatial domain, then each f_l is fit with effective ranges varying in the interval $\left(\frac{\delta_l}{5},\ \frac{\delta_1}{\delta_l} \cdot \frac{w}{5}\right)$. We find the splines are nearly linear, so estimates of f_l are very accurate even somewhat outside of the interval used for fitting.

Appendix C. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2022.107503.

References

Abdalati, W., Zwally, H.J., Bindschadler, R., Csatho, B., Farrell, S.L., Fricker, H.A., Harding, D., Kwok, R., Lefsky, M., Markus, T., et al., 2010. The ICESat-2 laser altimetry mission. Proc. IEEE 98, 735–751.

Berrocal, V.J., Raftery, A.E., Gneiting, T., 2008. Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. Ann. Appl. Stat. 2, 1170–1193.

Bradley, J.R., Cressie, N., Shi, T., 2013. Comparing and selecting spatial predictors using local criteria. Technical Report. Centre for Statistical and Survey Methodology, University of Wallongong. Working Paper 21-13.

Cook, B.D., Nelson, R.F., Middleton, E.M., Morton, D.C., McCorkel, J.T., Masek, J.G., Ranson, K.J., Ly, V., Montesano, P.M., et al., 2013. NASA Goddard's LiDAR, hyperspectral and thermal (g-liht) airborne imager. Remote Sens. 5, 4045–4066.

DHS Program, 2019. The DHS program – AIDS indicator surveys (AIS). https://dhsprogram.com/What-We-Do/Survey-Types/AIS.cfm.

Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., Armston, J., Tang, H., Duncanson, L., Hancock, S., Jantz, P., Marselis, S., Patterson, P.L., Qi, W., Silva, C., 2020. The global ecosystem dynamics investigation: high-resolution laser ranging of the Earth's forests and topography. Sci. Remote Sens. 1, 100002.

Filippone, M., Zhong, M., Girolami, M., 2013. A comparative evaluation of stochastic-based inference methods for Gaussian process models. Mach. Learn. 93, 93–114.

Finley, A., Datta, A., Banerjee, S., 2020a. spNNGP: spatial regression models for large datasets using nearest neighbor Gaussian processes. https://CRAN.R-project.org/package=spNNGP. r package version 0.1.3.

Finley, A.O., Datta, A., Banerjee, S., 2020b. spNNGP R package for nearest neighbor Gaussian process models. preprint. arXiv:2001.09111.

Finney, M.A., 1998. FARSITE, Fire Area Simulator–model development and evaluation. US Department of Agriculture, Forest Service, Rocky Mountain Research Station.

Fuglstad, G.A., Hem, I.G., Knight, A., Rue, H., Riebler, A., et al., 2020. Intuitive joint priors for variance parameters. Bayesian Anal. 15, 1109–1137.

Fuglstad, G.A., Simpson, D., Lindgren, F., Rue, H., 2019. Constructing priors that penalize the complexity of Gaussian random fields. J. Am. Stat. Assoc. 114, 445–452.

Gelfand, A.E., Diggle, P., Guttorp, P., Fuentes, M., 2010. Handbook of Spatial Statistics. CRC Press.

Geyer, C.J., Meeden, G.D., 2005. Fuzzy and randomized confidence intervals and p-values. Stat. Sci. 20, 358–366.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102, 359–378.

Gómez-Rubio, V., 2020. Bayesian Inference with INLA. CRC Press.

Handock, M., Stein, M., 1993. A Bayesian analysis of kriging. Technometrics 35, 403–410.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S., Goetz, S.J., Loveland, T.R., et al., 2013. High-resolution global maps of 21st-century forest cover change. Science 342, 850–853.

Heaton, M.J., Datta, A., Finley, A.O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R.B., Hammerling, D., Katzfuss, M., et al., 2019. A case study competition among methods for analyzing large spatial data. J. Agric. Biol. Environ. Stat. 24, 398–425.

Higdon, D., 1998. A process-convolution approach to modelling temperatures in the North Atlantic Ocean. Environ. Ecol. Stat. 5, 173–190.

Hurtt, G.C., Dubayah, R., Drake, J., Moorcroft, P.R., Pacala, S.W., Blair, J.B., Fearon, M.G., 2004. Beyond potential vegetation: combining LiDAR data and a height-structured model for carbon studies. Ecol. Appl. 14, 873–883.

Hyman, J.M., 1983. Accurate monotonicity preserving cubic interpolation. SIAM J. Sci. Stat. Comput. 4, 645–654.

ICF International, 2012. Demographic and Health Survey Sampling and Household Listing Manual. ICF International, Calverton, Maryland, USA.

Katzfuss, M., Guinness, J., 2020. A general framework for Vecchia approximations of Gaussian processes. Stat. Sci., in press. https://doi.org/10.1214/19-STS755.

Katzfuss, M., Guinness, J., Gong, W., Zilber, D., 2018. Vecchia approximations of Gaussian-process predictions. preprint. arXiv:1805.03309.

Kennedy, M.C., O'Hagan, A., 2000. Predicting the output from a complex computer code when fast approximations are available. Biometrika 87, 1–13.

Kenya National Bureau of Statistics, Ministry of Health/Kenya, National AIDS Control Council/Kenya, Kenya Medical Research Institute, and National Council For Population And Development/Kenya, 2015. Kenya Demographic and Health Survey 2014. Rockville, Maryland, USA. http://dhsprogram.com/pubs/pdf/FR308/FR308.pdf.

Klein, T., Randin, C., Körner, C., 2015. Water availability predicts forest canopy height at the global scale. Ecol. Lett. 18, 1311–1320.

Lantuéjoul, C., 2013. Geostatistical Simulation: Models and Algorithms. Springer Science & Business Media.

Li, Z.R., Hsiao, Y., Godwin, J., Martin, B.D., Wakefield, J., Clark, S.J., 2019. Changes in the spatial distribution of the under five mortality rate: small-area analysis of 122 DHS surveys in 262 subregions of 35 countries in Africa. PLoS One 14, e0210645.

Lindgren, F., Rue, H., 2015. Bayesian spatial modelling with R-INLA. J. Stat. Softw. 63.

Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic differential equation approach (with discussion). J. R. Stat. Soc., Ser. B 73, 423–498.

Liu, H., Hitchcock, D.B., Samadi, S.Z., 2019. Spatial and spatio-temporal analysis of precipitation data from South Carolina. In: Modern Statistical Methods for Spatial and Multivariate Data. Springer, pp. 31–50.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., Sain, S., 2015. A multiresolution Gaussian process model for the analysis of large spatial datasets. J. Comput. Graph. Stat. 24, 579–599.

Nychka, D., Hammerling, D., Krock, M., Wiens, A., 2018. Modeling and emulation of nonstationary Gaussian fields. Spat. Stat. 28, 21–38.

Nychka, D., Hammerling, D., Sain, S., Lenssen, N., 2016. LatticeKrig: multiresolution kriging based on Markov random fields. www.image.ucar.edu/LatticeKrig. r package version 6.4.

Paige, J., Fuglstad, G.A., Riebler, A., Wakefield, J., 2020. Design- and model-based approaches to small-area estimation in a low and middle income country context: comparisons and recommendations. J. Surv. Stat. Methodol.

Rue, H., Follestad, T., 2001. GMRFLib: a C-library for fast and exact simulation of Gaussian Markov random fields. Technical Report. SIS-2002-236.
Rue, H., Held, L., 2005. Gaussian Markov Random Fields. Monographs on Statistics and Applied Probability, vol. 104. Chapman & Hall/CRC, Boca Raton, FL.
Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). J. R. Stat. Soc., Ser. B 71, 319–392.
Rue, H., Riebler, A., Sørbye, S.H., Illian, J.B., Simpson, D.P., Lindgren, F.K., 2017. Bayesian computing with INLA: a review. Annu. Rev. Stat. Appl. 4, 395–421.
Schutz, B.E., Zwally, H.J., Shuman, C.A., Hancock, D., DiMarzio, J.P., 2005. Overview of the ICESat mission. Geophys. Res. Lett. 32.
Simpson, D., Rue, H., Riebler, A., Martins, T., Sørbye, S., 2017. Penalising model component complexity: a principled, practical approach to constructing priors (with discussion). Stat. Sci. 32, 1–28.
Sjöstedt-de Luna, S., Young, A., 2003. The bootstrap and kriging prediction intervals. Scand. J. Stat. 30, 175–192.
Stein, M., 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer.
Stevens, F.R., Gaughan, A.E., Linard, C., Tatem, A.J., 2015. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. PLoS ONE 10, e0107042.
Tatem, A.J., 2017. WorldPop, open data for spatial demography. Sci. Data 4.
The World Bank, 2019. Living standards measurement study (LSMS) | surveyunit. http://surveys.worldbank.org/lsms.
Thomas, Z.M., 2015. Bayesian hierarchical space-time clustering methods. Ph.D. thesis. The Ohio State University.
Thomas, Z.M., Matsuo, T., Nychka, D.W., Cousins, E.D., Wiltberger, M.J., 2014. Multi-resolution assimilative analysis of high-latitude ionospheric convection in both hemispheres. In: AGU Fall Meeting Abstracts, SA21A–4046.
Tuckwell, H.C., 2018. Elementary Applications of Probability Theory. Routledge.
UNICEF - Statistics and Monitoring, 2012. Multiple Indicator Cluster Surveys (MICS). http://www.unicef.org/statistics/index_24302.html.
United Nations, 2020. Sustainable development goals. http://sustainabledevelopment.un.org/owg.html.
USAID, 2019. Demographic and health surveys. United States Agency for International Development. http://www.dhsprogram.com.
Wagner, Z., Heft-Neal, S., Bhutta, Z.A., Black, R.E., Burke, M., Bendavid, E., 2018. Armed conflict and child mortality in Africa: a geospatial analysis. Lancet 392 (10150), 857–865.
Wakefield, J., Fuglstad, G.A., Riebler, A., Godwin, J., Wilson, K., Clark, S., 2019. Estimating under five mortality in space and time in a developing world context. Stat. Methods Med. Res. 28, 2614–2634.
Wendland, H., 1995. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. Adv. Comput. Math. 4, 389–396.
Zilber, D., Katzfuss, M., 2019. Vecchia-Laplace approximations of generalized Gaussian processes for big non-Gaussian spatial data. preprint. arXiv:1906.07828.