Supplementary material for "Constructing Priors that Penalize the Complexity of Gaussian Random Fields"

Geir-Arne Fuglstad[1], Daniel Simpson[2], Finn Lindgren[3], and Håvard Rue[4]

[1]Department of Mathematical Sciences, NTNU, Norway
[2]Department of Statistical Sciences, University of Toronto, Canada
[3]School of Mathematics, University of Edinburgh, United Kingdom
[4]CEMSE Division, King Abdullah University of Science and Technology, Saudi Arabia

December 6, 2017

S1 Proofs

S1.1 Theorem 2.1

Proof. For a fixed value of κ, the covariance matrix of $\boldsymbol{u} = (u(\boldsymbol{s}_1, u(\boldsymbol{s}_2), \ldots, u(\boldsymbol{s}_n))$ is $\Sigma(\tau) = \tau^2\Sigma_0$, where Σ_0 depends on the values of κ and ν and the locations at which the process is observed. This means that $\boldsymbol{u}|\tau, \kappa \sim \mathcal{N}_n(\boldsymbol{0}, \tau^2\Sigma_0)$, and we need to derive the PC prior for the scale parameter of a multivariate Gaussian distribution with the base model $\tau = 0$.

This can be formulated as constructing the PC prior for a precision parameter, which was done in Simpson et al. (2017, Appendix A.2), and a transformation to a scale parameter results in $\pi(\tau|\kappa) = \lambda \exp(-\lambda\tau)$, for $\tau > 0$, where $\lambda > 0$. □

S1.2 Theorem 2.2

Proof. The marginal standard deviation is given by $\sigma = \tau\kappa^{-\nu}C^{-1}$, where

$$C = \sqrt{\frac{\Gamma(\nu + d/2)(4\pi)^{d/2}}{\Gamma(\nu)}}.$$

The probability $\mathrm{P}(\sigma > \sigma_0|\kappa) = \alpha$ is equivalent to $\mathrm{P}(\tau > \sigma_0\kappa^\nu C|\kappa) = \alpha$ and under the prior distribution this leads to

$$\exp(-\lambda\sigma_0\kappa^\nu C) = \alpha$$
$$\lambda = -\kappa^{-\nu}\sqrt{\frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{d/2}}}\frac{\log(\alpha)}{\sigma_0}.$$

□

S1.3 Theorem 2.3

Proof. Restrict u to the subset $[0, L]^d \subset \mathbb{R}^d$ and let $\kappa = \kappa_0 > 0$ denote the base model. Let $1/L = o(\kappa_0)$, then the covariance function on $[0, L]^d$ is, for small κ_0, well approximated by

$$c(\boldsymbol{s}, \boldsymbol{t}) = \sum_{\boldsymbol{w} \in \frac{2\pi}{L}\mathbb{Z}^d} f_\nu(\boldsymbol{w}; \kappa, \tau) \exp\left(-\imath \langle \boldsymbol{w}, \boldsymbol{s} - \boldsymbol{t} \rangle\right) \left(\frac{2\pi}{L}\right)^d,$$

where $f_\nu(\boldsymbol{w}; \kappa, \tau) = (2\pi)^{-d} \tau^2 (\kappa^2 + ||\boldsymbol{w}||^2)^{-(\nu + d/2)}$ is the spectral density of a Matérn GRF on $\mathbb{R}^d$ with parameters τ, κ and ν (See Lindgren et al. (2011)). Further, the KLD for the periodic approximation from κ_0 to κ, for a fixed τ, is (based on Bogachev (1998, Thm. 6.4.6))

$$\begin{aligned}
\mathrm{KL}(\kappa, \kappa_0) &= \frac{1}{2} \sum_{\boldsymbol{w} \in \frac{2\pi}{L}\mathbb{Z}^d} \left[\frac{f_\nu(\boldsymbol{w}; \kappa, \tau)}{f_\nu(\boldsymbol{w}; \kappa_0, \tau)} - 1 - \log \frac{f_\nu(\boldsymbol{w}; \kappa, \tau)}{f_\nu(\boldsymbol{w}; \kappa_0, \tau)} \right] \\
&= \frac{1}{2} \sum_{\boldsymbol{w} \in \frac{2\pi}{L}\mathbb{Z}^d} \left[\frac{(\kappa_0^2 + ||\boldsymbol{w}||^2)^\alpha}{(\kappa^2 + ||\boldsymbol{w}||^2)^\alpha} - 1 - \log \frac{(\kappa_0^2 + ||\boldsymbol{w}||^2)^\alpha}{(\kappa^2 + ||\boldsymbol{w}||^2)^\alpha} \right],
\end{aligned}$$

where $\alpha = \nu + d/2$.

The sum can be divided in two parts: the zero frequency E_0 and the other frequencies E_1. The zero frequency term is

$$E_0 = \frac{1}{2} \left[\left(\frac{\kappa_0^2}{\kappa^2}\right)^\alpha - 1 - \log \left(\frac{\kappa_0^2}{\kappa^2}\right)^\alpha \right]$$

and the remaining terms are

$$\begin{aligned}
E_1 &= \frac{1}{2} \left(\frac{L\kappa}{2\pi}\right)^d \sum_{\boldsymbol{w} \in \frac{2\pi}{L\kappa}\mathbb{Z}^d, \boldsymbol{w} \neq 0} \left[\frac{((\kappa_0/\kappa)^2 + ||\boldsymbol{w}||^2)^\alpha}{(1 + ||\boldsymbol{w}||^2)^\alpha} - 1 - \log \frac{((\kappa_0/\kappa)^2 + ||\boldsymbol{w}||^2)^\alpha}{(1 + ||\boldsymbol{w}||^2)^\alpha} \right] \left(\frac{2\pi}{L\kappa}\right)^d \\
&= \frac{1}{2} \left(\frac{L\kappa}{2\pi}\right)^d \left\{ \int_{\mathbb{R}^d} \left[\frac{((\kappa_0/\kappa)^2 + ||\boldsymbol{w}||^2)^\alpha}{(1 + ||\boldsymbol{w}||^2)^\alpha} - 1 - \log \frac{((\kappa_0/\kappa)^2 + ||\boldsymbol{w}||^2)^\alpha}{(1 + ||\boldsymbol{w}||^2)^\alpha} \right] \mathrm{d}\boldsymbol{w} + o(1) \right\}
\end{aligned}$$

as $\kappa_0 \to 0$. For any fixed value of L, we may include $(L/(2\pi))^d$ in the hyperparameter λ in the PC prior. Thus we consider the rescaled terms

$$\tilde{E}_0 = \frac{1}{2} \left(\frac{2\pi}{L}\right)^d \left[\left(\frac{\kappa_0^2}{\kappa}\right)^\alpha - 1 - \log \left(\frac{\kappa_0^2}{\kappa}\right)^\alpha \right]$$

and

$$\tilde{E}_1 = \frac{1}{2} \kappa^d \left\{ \int_{\mathbb{R}^d} \left[\frac{((\kappa_0/\kappa)^2 + ||\boldsymbol{w}||^2)^\alpha}{(1 + ||\boldsymbol{w}||^2)^\alpha} - 1 - \log \frac{((\kappa_0/\kappa)^2 + ||\boldsymbol{w}||^2)^\alpha}{(1 + ||\boldsymbol{w}||^2)^\alpha} \right] \mathrm{d}\boldsymbol{w} + o(1) \right\}.$$

Let $\kappa_0 \to 0$ with $1/L = o(\kappa_0)$, then $\tilde{E}_0 \to 0$ and, for $d \leq 3$,

$$\tilde{E}_1 \to \frac{1}{2}\kappa^d \int_{\mathbb{R}^d} \left[\left(\frac{||\boldsymbol{w}||^2}{(1+||\boldsymbol{w}||^2)} \right)^\alpha - 1 - \log\left(\frac{||\boldsymbol{w}||^2}{(1+||\boldsymbol{w}||^2)} \right)^\alpha \right] \mathrm{d}\boldsymbol{w} = \frac{1}{2} C_\alpha \kappa^d,$$

where the finiteness of the integral is shown to hold in Section S8 of the Supplementary Material, and C_α is a constant that depends on α and d.

The distance from the base model is $\tilde{\text{dist}}(\kappa) = \sqrt{2 \cdot \frac{1}{2} C_\alpha \kappa^d}$, where we can absorb the constants in the hyperparameter of the PC prior, so we choose the distance $\text{dist}(\kappa) = \kappa^{d/2}$. The PC prior using an exponential distribution on the distance is then

$$\pi(\kappa) = \lambda \exp(-\lambda \kappa^{d/2}) \frac{\mathrm{d}}{\mathrm{d}\kappa} \kappa^{d/2} = \frac{d\lambda}{2} \kappa^{d/2-1} \exp(-\lambda \kappa^{d/2}), \quad \kappa > 0.$$

□

S1.4 Theorem 2.4

Proof. The probability $\mathrm{P}(\rho < \rho_0) = \alpha$ is equivalent to $\mathrm{P}(\kappa > \sqrt{8\nu}/\rho_0) = \alpha$ and under the prior distribution we find

$$\begin{aligned} \exp(-\lambda(\sqrt{8\nu}/\rho_0)^{d/2}) &= \log(\alpha) \\ \lambda &= -\left(\frac{\rho_0}{\sqrt{8\nu}} \right)^{d/2} \log(\alpha). \end{aligned}$$

□

S1.5 Theorem 2.5

Proof. Using Theorems 2.1 and 2.3, we find the joint prior

$$\begin{aligned} \pi(\kappa, \tau) &= \pi(\kappa)\pi(\tau|\kappa) \\ &= \frac{d}{2} \lambda_1 \kappa^{d/2-1} \exp(-\lambda_1 \kappa^{d/2}) \lambda_2 \exp(-\lambda_2 \tau) \\ &= \frac{d}{2} \lambda_1 \lambda_2 \kappa^{d/2-1} \exp(-\lambda_1 \kappa^{d/2} - \lambda_2 \tau), \quad \tau > 0, \kappa > 0. \end{aligned}$$

And Theorems 2.2 and 2.4 gives

$$\lambda_1 = -\left(\frac{\rho_0}{\sqrt{8\nu}} \right)^{d/2} \log(\alpha_1) \quad \text{and} \quad \lambda_2 = -\kappa^{-\nu} \sqrt{\frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{d/2}}} \frac{\log(\alpha_2)}{\sigma_0}.$$

□

S1.6 Theorem 2.6

Proof. Since there is no dependence of ρ on τ, the change of variables from (κ, τ) to (ρ, σ) can be divided in two steps. First, $\rho = \sqrt{8\nu}/\kappa$ so we find

$$\begin{aligned}\pi(\rho) &= \pi(\kappa = \sqrt{8\nu}/\rho) \left| \frac{\mathrm{d}}{\mathrm{d}\rho} \sqrt{8\nu}\rho^{-1} \right| \\ &= \frac{d}{2}\lambda_1 (\sqrt{8\nu})^{d/2-1} \rho^{-d/2+1} \exp(-\lambda_1 (\sqrt{8\nu})^{d/2} \rho^{-d/2}) \sqrt{8\nu}\rho^{-2} \\ &= \frac{d}{2}\tilde{\lambda}_1 \rho^{-d/2-1} \exp(-\tilde{\lambda}_1 \rho^{-d/2}), \quad \rho > 0,\end{aligned}$$

where $\tilde{\lambda}_1 = -\log(\alpha_1)\rho_0^{d/2}$.

Second, $\sigma = \tau\kappa^{-\nu}C^{-1}$, where

$$C = \sqrt{\frac{\Gamma(\nu + d/2)(4\pi)^{d/2}}{\Gamma(\nu)}}.$$

Note that conditioning on κ is equivalent to conditioning on ρ. So the density $\pi(\sigma|\rho)$ can be found by

$$\begin{aligned}\pi(\sigma|\rho) = \pi(\sigma|\kappa) &= \pi(\tau = \sigma\kappa^\nu C|\kappa) \left| \frac{\partial}{\partial\sigma} \sigma\kappa^\nu C \right| \\ &= \lambda_2 \exp(-\lambda_2 \sigma\kappa^\nu C)\kappa^\nu C \\ &= \tilde{\lambda}_2 \exp(-\tilde{\lambda}_2\sigma), \quad \sigma > 0,\end{aligned}$$

where $\tilde{\lambda}_2 = -\frac{\log(\alpha_2)}{\sigma_0}$. So the joint density is

$$\pi(\rho, \sigma) = \pi(\rho)\pi(\sigma|\rho) = \frac{d}{2}\tilde{\lambda}_1\tilde{\lambda}_2\rho^{-d/2-1} \exp(-\tilde{\lambda}_1\rho^{-d/2} - \tilde{\lambda}_2\sigma), \quad \rho > 0, \sigma > 0.$$

□

S2 Detailed discussion for bounded domains

The derivations in the main paper use the assumption that the size of the domain is large compared to the range of the base model. This is a reasonable assumption if the underlying GRF exists on a larger domain than the area on which observations have been made since the prior should be based on the distribution of the GRF on the domain where it is defined. However, if it is known that the GRF only exists on a bounded domain, it would be reasonable to base the derivation instead on the bounded domain and not a larger ambient domain.

When the parameter κ increases, the variance of the process increases, but the spread of the observations relative to each other does not change. Since there is no larger ambient

space in which this effect could be distinguished from adding an intercept to the model, it is more meaningful to have a base model with finite range.

When the priors are derived based on bounded domains, there will typically not be any analytic expressions available. One exception is the exponential covariance function on the one-dimensional domain $[0, L]$ where the exact expression for the distance between the models specified by (κ, τ) and (κ_0, τ) can be derived (see Section S3) and is given by

$$\text{dist}_{\text{1D,exp}}(\kappa||\kappa_0) = \sqrt{\frac{\kappa_0}{\kappa} - 1 - \log\left(\frac{\kappa_0}{\kappa}\right) + L\left(\frac{\kappa_0^2}{2\kappa} - \kappa_0 + \frac{\kappa}{2}\right)}. \tag{1}$$

In this case it is clear that the term $\log(\kappa)$ dominates when κ is small if $\sqrt{8\nu}/\kappa_0$ is of the same order as L or larger. In the following example one can see how the prior for κ calculated based on this distance differs from the one derived in the previous section

Example S2.1 (One-dimensional exponential covariance function)**.** Let a GRF with an exponential covariance function be observed on $[0, 1]$. There is a one-to-one correspondence between κ and ρ, so the distance given in Equation (1) can be expressed in range by

$$\text{dist}_{\rho_0}(\rho) = \sqrt{\frac{\rho}{\rho_0} - 1 - \log\left(\frac{\rho}{\rho_0}\right) + \sqrt{8\nu}\left(\frac{\rho}{2\rho_0^2} - \frac{1}{\rho_0} + \frac{1}{2\rho}\right)},$$

where ρ_0 is the range of the base model. This distance results in the prior

$$\pi_1(\rho) = \lambda_1 \exp(-\lambda_2 \text{dist}_{\rho_0}(\rho)) \frac{1}{2\text{dist}_{\rho_0}(\rho)} \left(\frac{1}{\rho} - \frac{1}{\rho_0} + \frac{\sqrt{8\nu}}{2}\left[\frac{1}{\rho^2} - \frac{1}{\rho_0^2}\right]\right), \quad 0 < \rho < \rho_0,$$

where $\lambda_1 = -\log(\alpha)/\text{dist}_{\rho_0}(R_0)$ ensures that $\text{P}(\rho < R_0) = \alpha$.

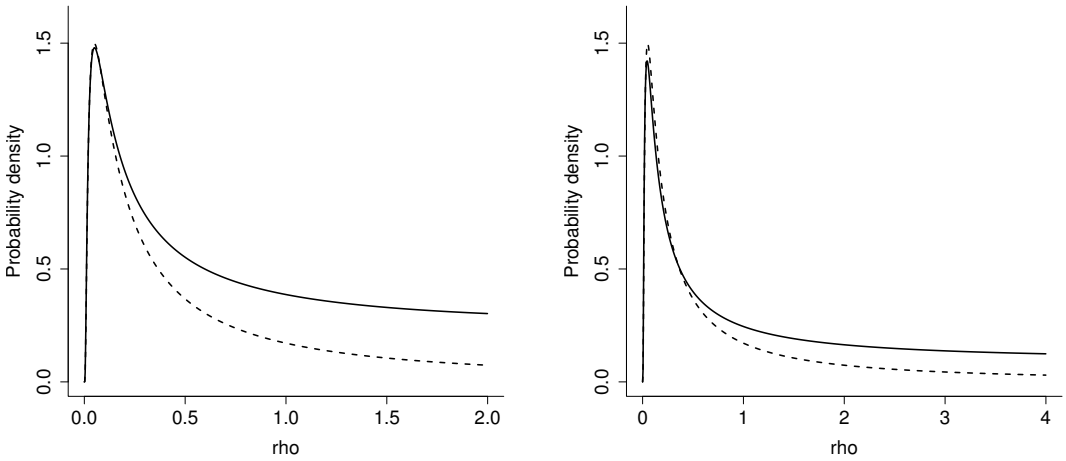
For comparison, the prior for ρ based on an unbounded domain is

$$\pi_2(\rho) = \frac{\lambda_2}{2}\rho^{-3/2}\exp(-\lambda_2\rho^{-1/2}), \quad \rho > 0,$$

where $\lambda_2 = -\log(\alpha)R_0^{1/2}$ ensures that $\text{P}(\rho < R_0) = \alpha$.

The parameter values $R_0 = 0.05$ and $\alpha = 0.05$ are chosen, and the prior based on the unbounded domain and the priors based on the bounded domain for $\rho_0 = 2$ and $\rho_0 = 4$ are shown in Figure S1. The figure shows that the two prior constructions are similar for $\rho < 0.25$ for $\rho = 2$ and for $\rho < 0.5$ for $\rho_0 = 4$. For higher values of range, π_2 distributes the probability mass over the entire positive line and has a faster decay than π_1. The priors will not correspond to each other in the case that $\rho_0 \to \infty$. In that case π_1 will have a decay of approximately $1/\rho$ whereas π_2 has a decay of approximately $\rho^{-3/2}$.

For other covariance functions and for bounded domains of dimension 2 and 3, the analytic expressions for the distances are not known, but the priors can be approximated numerically. The values of the prior on an interval $\kappa \in [A, B]$ can be computed by selecting a grid of sufficiently dense locations in the domain and then calculating the KLD



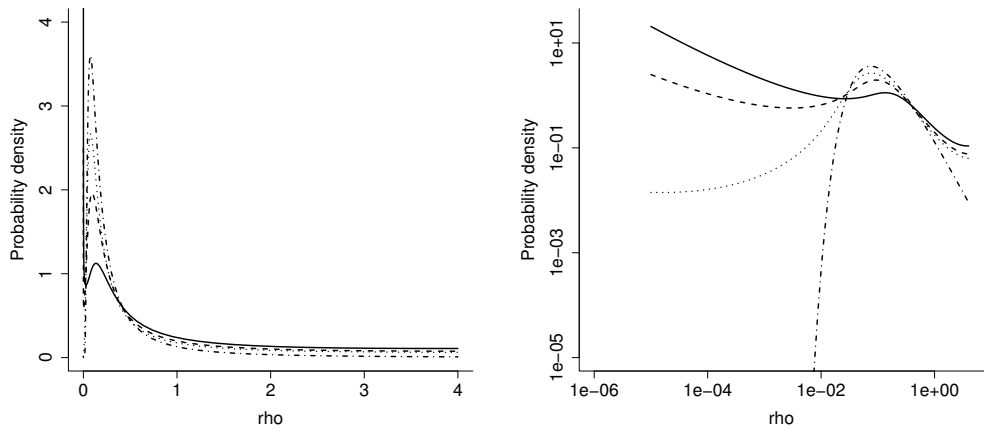
(a) Base model $\rho_0 = 2$ (b) Base model $\rho_0 = 4$

Figure S1: The PC prior based on the unbounded domain is shown as dashed lines in each subfigure and the prior based on the bounded domain is shown as solid lines for base model (a) $\rho_0 = 2$ (b) $\rho_0 = 4$. The prior based on the unbounded domain continues to $\rho = \infty$.

based on this finite set of locations. It would be possible to use a fully design-dependent prior where the KLD is calculated based on the observation locations in the dataset of interest, but this provides, potentially, undesired behaviour. Even for a bounded domain one is interested in doing predictions outside the observed locations on a much higher resolution and in that case using a prior that ignores all properties of the higher resolution process would not be advisable. If one wants to be able to do predictions at arbitrarily high resolution, the prior should be constructed based on the infinite-dimensional GRF defined on the full bounded domain. The following example demonstrates how calculations may be done in the two-dimensional case and the consequence of using a too low resolution in the calculation of the prior.

Example S2.2 (Two-dimensional Matérn covariance function with $\nu = 3/2$)**.** Let a GRF with a Matérn covariance function with smoothness $\nu = 1.5$ be observed on $[0, 1]^2$, and let the base model be given by $\rho_0 = 4$. We calculate priors π_1, π_2 and π_3 for the range based on a regular grid of 10×10 points, 20×20 points and 40×40 points, respectively, and prior π_4, which is the prior calculated based on an unbounded domain. For each prior the hyperparameter is set such that $P(\rho < 0.05) = 0.05$.

The calculated priors are shown in Figure S2 and demonstrate that the lower tail behaviour varies strongly dependent on the number of locations used to calculate the PC prior. One can see that the lower the resolution is, the higher the values of the prior in the left-hand side tail. Intuitively, this is because the distance decreases more slowly



(a) Original axes

(b) Log-log axes

Figure S2: PC priors for range in a Matérn GRF observed on $[0, 1]^2$ with smoothness $\nu = 3/2$. The prior based on the unbounded domain is shown as dash-dot-dashed line, and the priors for the bounded domain based on 10×10 points, 20×20 points and 40×40 are shown as a solid line, a dashed line and a dotted line, respectively. For each of the priors based on bounded domains the base model is $\rho_0 = 4$.

as a function of range as range goes to zero the lower the resolution is, and this causes more probability mass to be placed further out in the tail. This is because most of the differences in the GRFs for low ranges cannot be detected when it is observed at lower resolution. However, since the properties of the prior are used when making predictions at higher resolutions, we suggest to use the prior based on the continuous process instead of the discrete observation process.

As the resolution of the grid used to calculate the prior using the bounded domain goes to infinity, the prior will agree better and better with the prior based on the unbounded domain for low values of the range. In Example S2.1 the overlap for low values of range is clear since the analytic expression for infinite resolution is used.

Even though there are cases in which the domain of the GRF is naturally limited to the area on which the observations are made, the use of priors based on an unbounded domain have many advantages. The priors based on the bounded domain are more expensive to calculate and require the additional choice of a range for the base model, but have the same lower tail behaviour as the prior based on the unbounded domain and only the behaviour for higher ranges changes. Overall, we find that the unbounded domain prior is most appealing.

S3 Distance for exponential covariance function on bounded one-dimensional domain

S3.1 Goal

Let u_κ be a stationary GRF with the exponential covariance function,

$$c(d) = \frac{1}{2\kappa} \mathrm{e}^{-\kappa d}, \tag{2}$$

where $\kappa > 0$. This way of writing the exponential covariance function differs from the traditional parametrization using the range and the marginal variance, and is chosen because the KLD between the distributions described by different values $\kappa > 0$ is finite. The parametrization describes how to move in the parameter space while keeping the KLD finite. The goal of this appendix is to calculate the KLD between the distributions of u_κ and u_{κ_0} on the interval $[0, L]$

S3.2 Discretization

The direct computations for the interval $[0, L]$ are difficult. So we first consider the KLD for the distributions of u_κ and u_{κ_0} at the observation points $t_i = i\Delta t$, for $i = 0, 1, \ldots, N$, where $\Delta t = L/N$. The spatial field u_κ can be described as a stationary solution of the stochastic differential equation

$$\mathrm{d}u_\kappa(t) = -\kappa u_\kappa(t)\mathrm{d}t + \mathrm{d}W(t),$$

where W is a standard Wiener processes, and written explicitly as

$$u_\kappa(t) = \int_{-\infty}^{t} \mathrm{e}^{-\kappa(t-s)}\, \mathrm{d}W(s).$$

This expression shows that

$$u_\kappa(t_{i+1}) | u_\kappa(t_i) \sim \mathcal{N}(\mathrm{e}^{-\kappa\Delta t} u_\kappa(t_i), \sigma_\kappa^2),$$

where

$$\sigma_\kappa^2 = \mathrm{Var}[u_\kappa(t+\Delta t)|u_\kappa(t)] = \int_t^{t+\Delta t} \mathrm{e}^{-2\kappa(t+\Delta t-s)}\, \mathrm{d}s = \frac{1-\mathrm{e}^{-2\kappa\Delta t}}{2\kappa}.$$

This is an AR(1) process with initial condition $u_\kappa(t_0) \sim \mathcal{N}(0, (2\kappa)^{-1})$, which means that $\boldsymbol{u}_\kappa = (u_\kappa(t_0), \ldots, u_\kappa(t_N))$ has a multivariate Gaussian distribution with mean $\mathbf{0}$ and precision matrix

$$\mathbf{Q}_\kappa = \frac{1}{\sigma_\kappa^2} \begin{bmatrix} 1 & -\mathrm{e}^{-\kappa\Delta t} & & & \\ -\mathrm{e}^{-\kappa\Delta t} & 1+\mathrm{e}^{-2\kappa\Delta t} & -\mathrm{e}^{-\kappa\Delta t} & & \\ & \ddots & \ddots & \ddots & \\ & & -\mathrm{e}^{-\kappa\Delta t} & 1+\mathrm{e}^{-2\kappa\Delta t} & -\mathrm{e}^{-\kappa\Delta t} \\ & & & -\mathrm{e}^{-\kappa\Delta t} & 1 \end{bmatrix}. \tag{3}$$

S3.3 Kullback-Leibler divergence

The vectors $\boldsymbol{u}_{\kappa_0}$ and $\boldsymbol{u}_{\kappa}$ have multivariate Gaussian distributions and the KLD from the distribution described by κ_0 to the distribution described by κ is

$$\mathrm{KL}(\kappa, \kappa_0) = \frac{1}{2}\left[\mathrm{tr}(\mathbf{Q}_{\kappa_0}\mathbf{Q}_{\kappa}^{-1}) - (N+1) - \log\left(\frac{|\mathbf{Q}_{\kappa_0}|}{|\mathbf{Q}_{\kappa}|}\right)\right].$$

We are interested in taking the limit $\Delta t \to 0$ to find the value corresponding to the KLD from u_{κ_0} to u_{κ}. This is done in two steps: first we consider the trace and the $N+1$ term, and then we consider the log-determinant term.

S3.3.1 Step 1

Let $f_{\kappa} = 1/\sigma_{\kappa}^2$, then the trace term can be written as

$$\begin{aligned}
&\mathrm{tr}(\mathbf{Q}_{\kappa_0}\Sigma_{\kappa}) \\
&\quad = f_{\kappa_0}\left[2c_{\kappa}(0) + \sum_{i=1}^{N-1}(1+\mathrm{e}^{-2\kappa_0\Delta t})c_{\kappa}(0) - 2\sum_{i=1}^{N}\mathrm{e}^{-\kappa_0\Delta t}c_{\kappa}(\Delta t)\right] \\
&\quad = f_{\kappa_0}\left[2c_{\kappa}(0) + (N-1)(1+\mathrm{e}^{-2\kappa_0\Delta t})c_{\kappa}(0) - 2N\mathrm{e}^{-\kappa_0\Delta t}c_{\kappa}(\Delta t)\right].
\end{aligned}$$

We extract the first summand and parts of the last summand, and combine with 2 from the $N+1$ term, to find the limit

$$\begin{aligned}
2f_{\kappa_0}[c_{\kappa}(0) - \mathrm{e}^{-\kappa_0\Delta t}c\kappa(\Delta t)] - 2 &= 2f_{\kappa_0}\frac{1-\mathrm{e}^{-(\kappa+\kappa_0)\Delta t}}{2\kappa} - 2 \\
&= \frac{\kappa+\kappa_0}{\kappa}\frac{f_{\kappa_0}/\Delta t}{f_{\kappa+\kappa_0}/\Delta t} - 2 \\
&\to \frac{\kappa_0-\kappa}{\kappa}.
\end{aligned}$$

For the remaining summands and the remaining $N-1$ from the $N+1$ term, we can simplify the expression as

$$\begin{aligned}
&S_3(\Delta t)\\
&\quad= (N-1)f_{\kappa_0}\left[\left(1+\mathrm{e}^{-2\kappa_0\Delta t}\right)c_\kappa(0) - 2\mathrm{e}^{-\kappa_0\Delta t}c_\kappa(\Delta t)\right] - (N-1)\\
&\quad= (N-1)f_{\kappa_0}\left[\left(1+\mathrm{e}^{-2\kappa_0\Delta t}\right)\frac{1}{2\kappa} - 2\frac{\mathrm{e}^{-(\kappa_0+\kappa)\Delta t}}{2\kappa}\right] - (N-1)\\
&\quad= (N-1)f_{\kappa_0}\frac{1}{2\kappa}\Bigg[1 + (1 - 2\kappa_0\Delta t + \frac{4\kappa_0^2(\Delta t)^2}{2})\\
&\qquad\quad - 2(1-(\kappa_0+\kappa)\Delta t + \frac{(\kappa_0+\kappa)^2(\Delta t)^2}{2}) + o((\Delta t)^2)\Bigg] - (N-1)\\
&\quad= (N-1)f_{\kappa_0}\frac{1}{2\kappa}\Bigg[(-2\kappa_0 + 2(\kappa_0+\kappa))\Delta t\\
&\qquad\quad + (2\kappa_0^2 - (\kappa_0+\kappa)^2)(\Delta t)^2 + o((\Delta t)^2)\Bigg] - (N-1)\\
&\quad= (N-1)f_{\kappa_0}\left[\Delta t + \frac{2\kappa_0^2 - (\kappa_0+\kappa)^2}{2\kappa}(\Delta t)^2 + o((\Delta t)^2)\right] - (N-1)\\
&\quad= \left(\frac{L}{\Delta t} - 1\right)\left(\frac{1}{\Delta t} + \kappa_0 + o(1)\right)\Bigg[\Delta t\\
&\qquad\quad + \frac{2\kappa_0^2 - (\kappa_0+\kappa)^2}{2\kappa}(\Delta t)^2 + o((\Delta t)^2)\Bigg] - \left(\frac{L}{\Delta t} - 1\right),
\end{aligned}$$

and see that the products involving $o(1)$ tend to zero

$$\begin{aligned}
S_3(\Delta t) &= L\left[\frac{1}{\Delta t} + \frac{2\kappa_0^2 - (\kappa_0+\kappa)^2}{2\kappa} - \frac{1}{\Delta t}\right] + L\kappa_0 - [1+o(1)] + 1\\
&= L\frac{4\kappa_0^2 - (\kappa_0+\kappa)^2}{2\kappa} + L\kappa_0 + o(1)\\
&= L\left(\kappa_0 + \frac{\kappa_0^2}{2\kappa} - \kappa_0 - \frac{\kappa}{2}\right) + o(1).
\end{aligned}$$

Thus the limit is

$$\mathrm{tr}(\mathbf{Q}_{\kappa_0}\Sigma_\kappa) - (N+1) \to \frac{\kappa_0}{\kappa} - 1 + L\left(\frac{\kappa_0^2}{2\kappa} - \frac{\kappa}{2}\right).$$

S3.3.2 Step 2

The determinant of the matrix in Equation (3) can be found by summing rows upwards, and we see that

$$|\mathrm{Q}| = \sigma^{-2(N+1)}(1 - \mathrm{e}^{-2\kappa\Delta t}) = 2\kappa\sigma^{-2N}.$$

Note that in the limit $\kappa \to 0$, $f \to \Delta t$ so the determinant behaves asymptotically as κ. This means that

$$\begin{aligned}\log\left(\frac{|\mathbf{Q}_{\kappa_0}|}{|\mathbf{Q}_{\kappa}|}\right) &= \log\left(\frac{2\kappa_0 f_{\kappa_0}^N}{2\kappa f_{\kappa}^N}\right) \\ &= \log\left(\frac{\kappa_0}{\kappa}\right) + N\log\left(\frac{f_{\kappa_0}}{f_{\kappa}}\right)\end{aligned}$$

and we need to find the limit of the second part,

$$\begin{aligned}&N\log\left(\frac{f_{\kappa_0}}{f_{\kappa}}\right) \\ &= \frac{L}{\Delta t}\left[\log\frac{1}{f_{\kappa}} - \log\frac{1}{f_{\kappa_0}}\right] \\ &= \frac{L}{\Delta t}\left[\log\left(\frac{1}{2\kappa}\left(1-\mathrm{e}^{-2\kappa\Delta t}\right)\right) - \log\left(\frac{1}{2\kappa_0}\left(1-\mathrm{e}^{-2\kappa_0\Delta t}\right)\right)\right] \\ &= \frac{L}{\Delta t}\left[\log\left(\Delta t - \kappa(\Delta t)^2 + o((\Delta t)^2)\right) - \log\left(\Delta t - \kappa_0(\Delta t)^2 + o((\Delta t)^2)\right)\right] \\ &= \frac{L}{\Delta t}\left[\log\left(1-\kappa\Delta t + o(\Delta t)\right) - \log\left(1-\kappa_0\Delta t + o(\Delta t)\right)\right] \\ &= \frac{L}{\Delta t}\left[-\kappa\Delta t + \kappa_0\Delta t + o(\Delta t)\right]\end{aligned}$$

Thus the limit is

$$\log\left(\frac{|\mathbf{Q}_{\kappa_0}|}{|\mathbf{Q}_{\kappa}|}\right) \to \log\left(\frac{\kappa_0}{\kappa}\right) + L(\kappa_0 - \kappa)$$

S3.4 Full KLD

The combination of the limits from the two steps gives the full KLD,

$$\begin{aligned}\mathrm{KL}(\kappa, \kappa_0) &= \frac{1}{2}\left[\frac{\kappa_0}{\kappa} - 1 + L\left(\frac{\kappa_0^2}{2\kappa} - \frac{\kappa}{2}\right) - \log\left(\frac{\kappa_0}{\kappa}\right) - L(\kappa_0 - \kappa)\right] \\ &= \frac{1}{2}\left[\frac{\kappa_0}{\kappa} - 1 - \log\left(\frac{\kappa_0}{\kappa}\right) + L\left(\frac{\kappa_0^2}{2\kappa} - \kappa_0 + \frac{\kappa}{2}\right)\right]. \qquad (4)\end{aligned}$$

S4 Details for the simulation study

In this section we present a small simulation study of the frequentist coverage of the credible intervals for the range and the marginal variance, and the behaviour of the joint posterior when using the PC prior, the Jeffreys' rule prior, and the Jeffreys prior for variance combined with a bounded uniform prior on range and a bounded uniform prior on the logarithm of range.

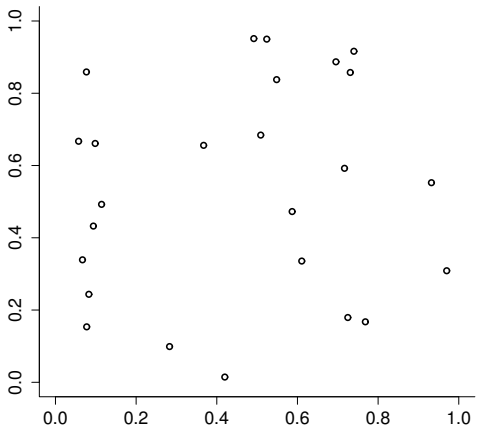


Figure S3: Spatial design for the simulation study.

S4.1 Study setup

We choose the observation domain $[0,1]^2 \subset \mathbb{R}^2$ and select the 25 observation locations, $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_{25}$, shown in Figure S3 at random. Then we simulate observations, $\boldsymbol{u} = (u(\boldsymbol{s}_1), u(\boldsymbol{s_2}), \ldots, u(\boldsymbol{s}_{25}))$, for these observation locations for a GRF with an exponential covariance function $c(r) = \exp(-2r/R_0)$ for $R_0 = 0.1$ and $R_0 = 1$. We generate 1000 realizations of the observations for $R_0 = 0.1$ and $R_0 = 1$ and collect them in datasets Data1 and Data2 , respectively. Additionally, for each of the 1000 realizations in Data1 a third dataset (Data3) is generated by simulating $y_i|p_i \sim \text{Binomial}(20, p_i)$, where $\text{probit}(p_i) = u_i$, for $i = 1, 2, \ldots, 25$.

Two models are used to fit the data: a spatial regression model (Model1) and a spatial logistic regression model (Model2). In Model1 observations are modelled as $y_i = u(\boldsymbol{s}_i)$, for $i = 1, 2, \ldots, 25$, where u is an exponential GRF with the covariance function, $c(r) = \sigma^2 \exp(-2r/\rho)$, where σ^2 is the marginal variance and ρ is the range, and in Model 2 the observations are modelled as $y_i|p_i \sim \text{Binomial}(20, p_i)$, where $\text{probit}(p_i) = u(\boldsymbol{s}_i)$, for $i = 1, 2, \ldots, 25$, where u is an exponential GRF with covariance function, $c(r) = \sigma^2 \exp(-2r/\rho)$, where σ^2 is the marginal variance and ρ is the range.

Four different priors are used for the parameters: the PC prior (PriorPC), the Jeffreys' rule prior (PriorJe), a uniform prior on range on a bounded interval combined with the Jeffreys' prior for variance (PriorUn1) and a uniform prior on the log-range on a bounded interval combined with the Jeffreys' prior for variance (PriorUn2). The full expressions for the priors are given in Table S1.

Table S1: The four different priors used in the simulation study. The Jeffreys' rule prior uses the spatial design of the problem through $U = (\frac{\partial}{\partial \rho}\Sigma)\Sigma^{-1}$, where Σ is the correlation matrix of the observations (See Berger et al. (2001)).

Prior	**Expression**	**Parameters**
PriorPC	$\pi_1(\rho,\sigma) = \lambda_1\lambda_2\rho^{-2}\exp\left(-\lambda_1\rho^{-1} - \lambda_2\sigma\right)$	$\rho, \sigma > 0$ Hyperparameters: α_ρ, ρ_0, α_σ, σ_0
PriorJe	$\pi_2(\rho,\sigma) = \sigma^{-1}\left(\text{tr}(U^2) - \frac{1}{n}\text{tr}(U)^2\right)^{1/2}$	$\rho, \sigma > 0$ Hyperparameters: None
PriorUn1	$\pi_3(\rho,\sigma) \propto \sigma^{-1}$	$\rho \in [A, B]$, $\sigma > 0$ Hyperparameters: A, B
PriorUn2	$\pi_4(\rho,\sigma) \propto \sigma^{-1}\cdot\rho^{-1}$	$\rho \in [A, B]$, $\sigma > 0$ Hyperparameters: A, B

S4.2 Frequentist coverage

The series of papers on reference priors for GRFs starting with Berger et al. (2001) evaluated the priors by studying frequentist properties of the resulting Bayesian inference. A prior intended for use as a default prior should lead to good frequentist properties such as frequentist coverage of the equal-tailed $100(1-\alpha)\%$ Bayesian credible intervals that is close to the nominal $100(1-\alpha)\%$. In this paper, the study is replicated with one key difference: no covariates are included. This choice is made because the PC prior is derived for a zero-mean GRF, and if a mean were desired, it would be handled by extending the hierarchical model with another latent component that had its own, separate prior. Without covariates the reference prior approach results in the Jeffreys' rule prior as there are no nuisance parameters to integrate out when constructing the spatial reference prior. Furthermore, we compute the $100(1-\alpha)\%$ highest posterior density (HPD) credible intervals (Chen and Shao, 1999) since the resulting posteriors will be highly skewed and the HPD intervals may differ substantially from the quantile-based intervals.

In this section Model1 is combined with PriorJe, PriorPC, PriorUn1 and PriorUn2. Data1 and Data2 each contains 1000 realizations and the frequentist coverage is estimated for the equal-tailed 95% credible intervals and the HPD 95% credible intervals for the range and the marginal variance by counting how many times the true parameter value is included in the credible intervals. The equal-tailed intervals are calculated based on the quantiles of the samples from an MCMC chain and the HPD intervals are calculated using the BOA package (Smith et al., 2007). We split the presentation of the results for the quantile-based credible intervals and the HPD credible intervals: in this section we discuss the quantile-based intervals and their associated results, and in the next section we discuss the results for the HPD credible intervals and differences from the results for the quantile-based intervals.

PriorJe has no hyperparameters, but PriorPC, PriorUn1 and PriorUn2 each has hyperparameters that need to be set before using the priors. For PriorUn1 and PriorUn2 it is hard to give guidelines about which values should be selected since the main purpose of limiting the prior distributions to a bounded interval is to avoid an improper posterior and the choice tends to be *ad-hoc*. For PriorPC, on the other hand, there is an interpretable statement for selecting the hyperparameters, which helps give an idea about which prior assumptions the chosen hyperparameters are expressing.

For PriorPC we need to make an *a priori* decision about the scales of the range and the marginal variance. The prior is set through four hyperparameters that describe our prior beliefs about the spatial field. We use $\mathrm{P}(\rho < \rho_0) = 0.05$ for $\rho_0 = 0.025\rho_{\mathrm{T}}$, $\rho_0 = 0.1\rho_{\mathrm{T}}$, $\rho_0 = 0.4\rho_{\mathrm{T}}$ and $\rho_0 = 1.6\rho_{\mathrm{T}}$, where ρ_{T} is the true range. This covers a prior where ρ_0 is much smaller than the true range, two priors where ρ_0 is smaller than the true range, but not far away, and one prior where ρ_0 is higher than the true range. For the marginal variance we use $\mathrm{P}(\sigma^2 > \sigma_0^2) = 0.05$, for $\sigma_0 = 0.625$, $\sigma_0 = 2.5$, $\sigma_0 = 10$ and $\sigma_0 = 40$. We follow the same logic as for range and cover too small and too large σ_0 and two reasonable values. For PriorUn1 and PriorUn2, we set the lower and upper limits for the nominal range according to the values $A = 0.05$, $A = 0.005$ and $A = 0.0005$, and

Table S2: Frequentist coverage of the 95% credible intervals for the range and the marginal variance when the true range is $\rho_T = 0.1$ using PriorPC. The average lengths of the credible intervals are shown in brackets.

(a) Range

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.0025	0.755 [0.24]	0.776 [0.22]	0.760 [0.20]	0.709 [0.18]
0.01	0.969 [0.33]	0.970 [0.32]	0.958 [0.28]	0.924 [0.21]
0.04	0.988 [0.46]	0.990 [0.41]	0.991 [0.33]	0.990 [0.25]
0.16	0.723 [0.99]	0.685 [0.82]	0.733 [0.55]	0.798 [0.34]

(b) Marginal variance

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.0025	0.960 [1.5]	0.943 [1.4]	0.946 [1.3]	0.898 [0.97]
0.01	0.934 [1.6]	0.966 [1.6]	0.960 [1.4]	0.923 [0.99]
0.04	0.949 [2.0]	0.945 [1.8]	0.953 [1.5]	0.941 [1.1]
0.16	0.895 [3.8]	0.905 [3.2]	0.947 [2.2]	0.977 [1.3]

$B = 2$, $B = 20$ and $B = 200$. Some of the values are intentionally extreme to see the effect of misspecification.

The results for PriorPC are given in Tables S2 and S3 for the true ranges $\rho_T = 0.1$ and $\rho_T = 1$, respectively, and the tables for PriorUn1 and PriorUn2 are given in Section S5. PriorJe resulted in 98.3% coverage with average length of the credible intervals of 0.78 for range and 96.7% coverage and average length of the credible intervals of 2.6 for marginal variance for $\rho_T = 0.1$, and 95.6% coverage with average length of the credible intervals of 376 for range and 95.6% coverage with average length of the credible intervals of 295 for variance for $\rho_T = 1$. The tables show that for PriorPC, PriorUn1 and PriorUn2 the coverage and the length of the credible intervals are sensitive to the choice of hyperparameters. The lengths of the credible intervals are, in general, more well-behaved for $\rho_T = 0.1$ than for $\rho_T = 1$ because there is more information about the range available in the domain when the range is shorter.

The results verifies the observations by Berger et al. (2001) that the inference is overly sensitive to the hyperparameters for PriorUn1. The coverage and the length of the credible intervals are strongly dependent on the upper limit of the prior. For PriorUn2 the coverage is good in both the short range and long range case, but the length of the credible intervals are sensitive to the upper limit of the prior. For PriorJe the coverage is good, but the credible intervals are excessively long and the prior is computationally expensive and only computationally feasible for a low number of observation locations. The average length of the credible intervals for $\rho_T = 1$ for marginal variance is 295, which imply unreasonably high standard deviations. The high standard deviations do not seem consistent with observations drawn with true marginal variance equal to 1.

Further, the results show that the coverage for PriorPC is stable when a too low

Table S3: Frequentist coverage of the 95% credible intervals for the range and the marginal variance when the true range is $\rho_T = 1$ using PriorPC. The average lengths of the credible intervals are shown in brackets.

(a) Range

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.025	0.957 [12]	0.947 [7.3]	0.921 [3.3]	0.782 [1.4]
0.1	0.977 [14]	0.967 [8.5]	0.962 [3.5]	0.861 [1.5]
0.4	0.963 [25]	0.970 [13]	0.988 [5.2]	0.980 [1.9]
1.6	0.63 [73]	0.301 [32]	0.711 [11]	0.945 [3.3]

(b) Marginal variance

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.025	0.956 [11]	0.949 [6.5]	0.927 [2.8]	0.771 [1.1]
0.1	0.964 [13]	0.966 [7.5]	0.950 [3.1]	0.848 [1.2]
0.4	0.953 [22]	0.964 [12]	0.980 [4.5]	0.965 [1.5]
1.6	0.435 [69]	0.549 [29]	0.804 [9.1]	0.988 [2.5]

lower limit for range or a too high upper limit for marginal variance is specified, but that specifying a too high lower limit for the range or a too low upper limit for variance produces large changes in the coverage. This is not unreasonable as the prior is then explicitly stating that the true value for range or variance is unlikely. The average length of the credible intervals are more sensitive to the hyperparameters than the coverages, but we see less extreme sizes for the credible intervals than for PriorJe.

With respect to computation time and ease of use versus coverage and length of credible intervals PriorUn2 and PriorPC appear to be the best choices. If coverage is the only concern, PriorUn2 performs the best, but if one also wants to control the length of the credible intervals by disallowing unreasonably high variances, PriorPC offers the most interpretable alternative. In a realistic situation it is highly likely that the researcher has prior knowledge, for example, that the spatial effect should not be greater than, say 4, and by encoding this information in PriorPC one can limit the upper limits of the credible intervals both for range and marginal variance.

S4.3 Differences in results between equal-tailed and HPD credible intervals

For each case discussed in the previous section, we also calculated the 95% credible intervals using HPD intervals and the results and tables are found in Section S6. In general, the average length of the credible intervals are significantly shorter for HPD credible intervals than for quantile-based credbile intervals, and the most substantial decrease is seen for true range equal to 1.0 using PriorJe, where the average length of the credible interval decreases from 376 to 95 for range and from 295 to 75 for marginal

variance. However, the conclusions in the previous section on differences in average lengths of credible intervals between priors and between true range equal to 0.1 and 1.0 remain valid since the relative differences remain similar. In particular, the average length of the HPD credible intervals for marginal variance for PriorJe with true range equal to 1.0 is around 95, which is still unreasonably high when prior knowledge about the marginal standard deviation is available.

The coverage of the credible intervals constructed using HPD intervals differ from the quantile-based intervals. For PriorPC, the coverage of the HPD intervals is more sensitive to the hyperparameters and if $\rho_0 = 0.4\rho_T$ the coverage of the HPD intervals for range are almost 100%. Further, the coverage of the HPD intervals for marginal variance are excessively high when $\sigma_0 = 40$ or $\sigma_0 = 10$, and there is no recommendation for hyperparameters that perform consistently well in both true range equal to 0.1 and 1.0 and for both range and marginal variance. Similarly, the coverage of the credible interval for range is almost 100% when the true range is 0.1 with PriorJe. This contrasts the quantile-based credible intervals where PriorJe performs well with respect to coverage for both true range equal to 0.1 and 1.0. For PriorUn1 and PriorUn2 the coverage of the HPD credible intervals are less sensitive to hyperparameters than the quantile-based credible intervals, but the the HPD intervals tend to have higher coverage than the nominal level.

We use the equal-tailed 95% credible intervals in what follows since the 95% HPD credible intervals are further away from nominal level and more sensitive to hyperparameters than equal-tailed 95% credible intervals for the PC prior.

S4.4 Behaviour of the joint posterior

In the previous section we only studied the marginal properties of the posterior, but these do not tell the entire story because there is strong dependence between the range and the marginal variance in the posterior distribution. We study this dependence using one realization from Data2 where the true range is 1 and the observed values lie in the range -1 to 3. An MCMC sampler is used to draw samples from the posterior of the marginal standard deviation and the range. Model1 is combined with PriorJe, and PriorPC with hyperparameters $\alpha_\rho = 0.05$, $\rho_0 = 0.1$, $\alpha_\sigma = 0.05$ and $\sigma_0 = 10$.

Figure S4 shows the strong posterior dependence between the marginal standard deviation and the range in the tail of the distribution. The long tails are not a major concern for predictions since the asymptotic predictions are the same along the ridge, but they pose a concern for the interpretability of the range and the marginal variance. Since the values of the observations lie within the range -1 to 3, it is unlikely that the true standard deviation should be on the order of 20.

As seen in Figure S5, the heavier upper tail for the joint posterior when using PriorJe compared to using PriorPC results in heavier tails also for the marginal posteriors. The lower endpoints of the equal-tailed credible intervals are similar using both priors, but there is a large difference in the upper endpoints. The PC prior for range has a heavy upper tail and the upper tail of the posterior for the range is controlled through the prior on the marginal variance. The light upper tail of the prior on marginal variance restricts the joint posterior from moving far along the ridge in the likelihood.

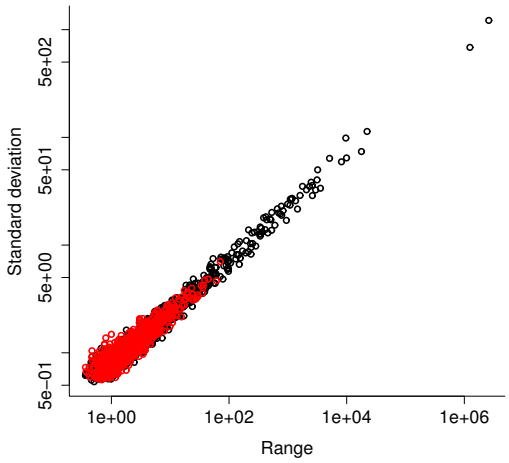
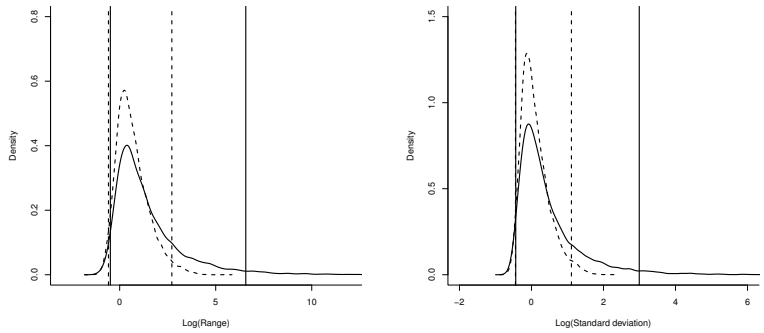


Figure S4: Samples from the joint posterior of range and marginal standard deviation. The red circles are samples using the PC-prior and the black circles are samples using the Jeffreys' rule prior.



(a) Posterior for the logarithm of range

(b) Posterior for the logarithm of marginal standard deviation

Figure S5: Marginal posteriors of the logarithms of range and marginal standard deviation. The dashed lines corresponds to the PC prior and the solid corresponds to the Jeffreys' rule prior. Equal-tailed 95% credible intervals are shown as vertical lines.

Intrinsic models have a place in statistics, but the results show that PriorJe favors intrinsic GRFs with large marginal standard deviations and ranges even though they might not be physically reasonable for the application. PriorPc offers a way to introduce prior belief about the size of the marginal standard deviations, and thus a way to reduce the preference for the intrinsic GRFs and limit the size of the credible intervals according to knowledge about the process.

S4.5 Example: Spatial logistic regression

A clear weakness of the reference priors is that they must be re-derived when components are added to the model or the observation process is changed. The PC prior can be used in any model since its derivation is observation-process agnostic and results in a prior for the model component itself and not the whole model. The frequentist coverage resulting from using Model2 with PriorPC for 500 of the realizations in Data3 is estimated similarly as in Section S4.2.

The experiment is repeated for 64 different settings of the prior: the hyperparameter ρ_0 varies over $\rho_0 = 0.0025, 0.01, 0.04, 0.16$ and the hyperparameter σ_0 varies over $\sigma_0 = 40, 10, 2.5, 0.625$. This covers a broad range of values from too small to too large. The values in Table S4 are similar to the values in Table S2 except that the equal-tailed credible intervals are slightly longer. The longer credible intervals are reasonable since the binomial likelihood gives less information about the spatial field than direct observations. The coverage for the marginal variance is good even for grossly miscalibrated priors, but the coverage for range is sensitive to bad calibration for range and the coverage is somewhat higher than nominal for the well-calibrated priors. This is a feature also seen in the directly observed case in Section S4.2. For completeness, the corresponding 95% HPD credible intervals are shown in Table S15. The table shows that the coverage for range is too low for $\rho_0 = 0.0025$ and $\rho_0 = 0.01$ and that the coverage is too high for $\rho_0 = 0.04$ and $\rho_0 = 0.16$. This was also the case for Gaussian observations, and the HPD intervals are more sensitivity to the hyperparameters than the equal-tailed credible intervals for PriorPC.

S5 Additional tables for simulation study using quantile-based credible intervals

The simulation study in Section S4 was run with four different priors: the PC prior (PriorPC), the Jeffreys' rule prior (PriorJe), a uniform prior on range on a bounded interval combined with the Jeffreys' prior for variance (PriorUn1) and a uniform prior on the log-range on a bounded interval combined with the Jeffreys' prior for variance (PriorUn2). For each prior a selection of hyperparameters were tested on datasets generated from true ranges $\rho_T = 0.1$ and $\rho_T = 1.0$, and the frequentist coverages of the 95% credible intervals and the lengths of the credible intervals were estimated. For $\rho_T = 0.1$, PriorJe gave 98.3% coverage with average length 0.78 for range and 96.7% coverage with average length 2.6 for marginal variance, and for $\rho_T = 1.0$, PriorJe gave 95.6% coverage with

Table S4: Frequentist coverage of the 95% credible intervals for range and marginal variance when the true range is 0.1 and true marginal variance is 1, where the average length of the credible intervals are given in brackets, for the spatial logistic regression example.

(a) Range

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.0025	0.790 [0.32]	0.775 [0.25]	0.760 [0.22]	0.720 [0.19]
0.01	0.982 [0.42]	0.981 [0.37]	0.974 [0.30]	0.960 [0.25]
0.04	0.990 [0.65]	0.987 [0.53]	0.995 [0.40]	0.985 [0.30]
0.16	0.621 [1.6]	0.638 [1.2]	0.682 [0.71]	0.779 [0.43]

(b) Marginal variance

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.0025	0.953 [2.1]	0.936 [1.9]	0.941 [1.7]	0.913 [1.2]
0.01	0.952 [2.2]	0.949 [2.1]	0.954 [1.7]	0.931 [1.2]
0.04	0.949 [2.7]	0.942 [2.5]	0.960 [1.9]	0.923 [1.3]
0.16	0.906 [5.5]	0.923 [4.2]	0.961 [2.7]	0.972 [1.5]

average length 376 for range and 95.6% coverage with average length of 295 for marginal variance. The results for PriorPC is given in Section S4 and the results for the two other priors are collected in the tables:

Prior	$\rho_T = \mathbf{0.1}$	$\rho_T = \mathbf{1.0}$
PriorUn1	Table S5	Table S7
PriorUn2	Table S6	Table S8

S6 Results of simulation study using HPD credible intervals

The simulation study in Section S4 was run with four different priors: the PC prior (PriorPC), the Jeffreys' rule prior (PriorJe), a uniform prior on range on a bounded interval combined with the Jeffreys' prior for variance (PriorUn1) and a uniform prior on the log-range on a bounded interval combined with the Jeffreys' prior for variance (PriorUn2). For each prior a selection of hyperparameters were tested on datasets generated from true ranges $\rho_T = 0.1$ and $\rho_T = 1.0$, and the frequentist coverages of the 95% highest posterior density (HPD) credible intervals and the average lengths of the HPD credible intervals were estimated. For $\rho_T = 0.1$, PriorJe gave 99.9% coverage with average length 0.46 for range and 98.2% coverage with average length 1.8 for marginal variance, and for $\rho_T = 1.0$, PriorJe gave 95.7% coverage with average length 95 for range and 96.5% coverage with average length of 75 for marginal variance. The results for PriorPC, PriorUn1 and PriorUn2 are given in the tables:

Table S5: Frequentist coverage of 95% credible intervals for range and marginal variance when the true range $\rho_T = 0.1$ using PriorUn1, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.920 [0.93]	0.886 [8.5]	0.840 [119]
$5 \cdot 10^{-3}$	0.937 [0.94]	0.910 [8.1]	0.866 [104]
$5 \cdot 10^{-4}$	0.937 [0.91]	0.925 [8.0]	0.864 [108]

(b) Marginal variance

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.941 [3.5]	0.937 [30]	0.900 [443]
$5 \cdot 10^{-3}$	0.934 [3.4]	0.924 [27]	0.924 [383]
$5 \cdot 10^{-4}$	0.934 [3.3]	0.945 [27]	0.922 [388]

Table S6: Frequentist coverage of 95% credible intervals for range and marginal variance when the true range $\rho_T = 0.1$ using PriorUn2, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.987 [0.44]	0.983 [0.72]	0.985 [1.1]
$5 \cdot 10^{-3}$	0.959 [0.43]	0.972 [0.74]	0.965 [1.3]
$5 \cdot 10^{-4}$	0.923 [0.39]	0.944 [0.68]	0.933 [1.1]

(b) Marginal variance

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.954 [1.9]	0.954 [2.7]	0.961 [3.5]
$5 \cdot 10^{-3}$	0.957 [1.7]	0.957 [2.4]	0.950 [3.8]
$5 \cdot 10^{-4}$	0.947 [1.6]	0.954 [2.3]	0.939 [3.2]

Table S7: Frequentist coverage of 95% credible intervals for range and marginal variance when the true range $\rho_T = 1$ using PriorUn1, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.995 [1.5]	0.840 [18]	0.562 [188]
$5 \cdot 10^{-3}$	0.997 [1.5]	0.831 [18]	0.560 [188]
$5 \cdot 10^{-4}$	0.993 [1.5]	0.823 [18]	0.550 [188]

(b) Marginal variance

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.975 [2.0]	0.848 [20]	0.574 [205]
$5 \cdot 10^{-3}$	0.978 [2.0]	0.822 [21]	0.600 [203]
$5 \cdot 10^{-4}$	0.983 [2.0]	0.837 [20]	0.564 [206]

Table S8: Frequentist coverage of 95% credible intervals for range and marginal variance when the true range $\rho_T = 1$ using PriorUn2, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.978 [1.5]	0.965 [13]	0.963 [69]
$5 \cdot 10^{-3}$	0.969 [1.5]	0.951 [12]	0.944 [67]
$5 \cdot 10^{-4}$	0.978 [1.5]	0.957 [13]	0.947 [68]

(b) Marginal variance

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.964 [1.8]	0.961 [12]	0.949 [61]
$5 \cdot 10^{-3}$	0.957 [1.8]	0.953 [11]	0.934 [60]
$5 \cdot 10^{-4}$	0.958 [1.8]	0.945 [12]	0.941 [59]

Table S9: Frequentist coverage of the 95% HPD credible intervals for the range and the marginal variance when the true range is $\rho_T = 0.1$ using PriorPC. The average lengths of the credible intervals are shown in brackets.

(a) Range

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.0025	0.571 [0.17]	0.586 [0.17]	0.584 [0.16]	0.535 [0.14]
0.01	0.903 [0.25]	0.912 [0.25]	0.900 [0.23]	0.841 [0.18]
0.04	1.000 [0.35]	0.999 [0.33]	0.999 [0.28]	0.998 [0.22]
0.16	0.990 [0.67]	0.992 [0.60]	0.980 [0.45]	0.957 [0.31]

(b) Marginal variance

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.0025	0.961 [1.3]	0.947 [1.3]	0.947 [1.2]	0.857 [0.92]
0.01	0.959 [1.4]	0.969 [1.4]	0.958 [1.2]	0.882 [0.93]
0.04	0.980 [1.7]	0.967 [1.6]	0.961 [1.3]	0.908 [1.0]
0.16	0.991 [2.8]	0.988 [2.5]	0.990 [1.9]	0.962 [1.2]

Prior	$\rho_T = \mathbf{0.1}$	$\rho_T = \mathbf{1.0}$
PriorPC	Table S9	Table S12
PriorUn1	Table S10	Table S13
PriorUn2	Table S11	Table S14
PriorPC and logistic regression	Table S15	N/A

S7 Prior for extra flexibility in the covariance structure

Lindgren et al. (2011) represented Matérn GRFs as the stationary solutions to the stochastic partial differential equation (SPDE)

$$[\kappa^2 - \Delta]^{\alpha/2}(\tau u(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathbb{R}^d, \tag{5}$$

where $\kappa > 0$ and $\tau > 0$ are parameters, α is connected to the smoothness ν through $\alpha = \nu + d/2$, Δ is the Laplacian, and $\mathcal{W}$ is standard Gaussian white noise. Ingebrigtsen et al. (2014) allowed the parameters of the SPDE to be spatially varying functions, $\log(\kappa(\cdot))$ and $\log(\tau(\cdot))$, through low-dimensional bases using a small number of covariates, and used independent Gaussian priors for the extra parameters. However, they experienced numerical problems and prior sensitivity, and Ingebrigtsen et al. (2015) developed an improved scheme for selecting the hyperparameters of the priors based on the properties of the resulting spatially varying local ranges and marginal variances. However, the inherent problem of their specification is that $\kappa(\cdot)$ affects both the correlation structure and the marginal variances of the spatial field. This makes it challenging to set priors on $\kappa(\cdot)$ and $\tau(\cdot)$, and we aim to improve their procedure by first improving the parametrization of the non-stationarity, and then developing a prior using the improved parametrization.

Table S10: Frequentist coverage of 95% HPD credible intervals for range and marginal variance when the true range $\rho_T = 0.1$ using PriorUn1, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.977 [0.71]	0.992 [5.7]	0.989 [92]
$5 \cdot 10^{-3}$	0.977 [0.74]	0.994 [5.6]	0.990 [78]
$5 \cdot 10^{-4}$	0.970 [0.71]	0.988 [5.4]	0.993 [82]

(b) Marginal variance

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.991 [2.7]	0.993 [19]	1.00 [312]
$5 \cdot 10^{-3}$	0.985 [2.7]	0.993 [18]	0.993 [263]
$5 \cdot 10^{-4}$	0.981 [2.6]	0.989 [18]	0.993 [270]

Table S11: Frequentist coverage of 95% HPD credible intervals for range and marginal variance when the true range $\rho_T = 0.1$ using PriorUn2, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.998 [0.34]	0.999 [0.45]	1.000 [0.54]
$5 \cdot 10^{-3}$	0.922 [0.33]	0.936 [0.46]	0.922 [0.62]
$5 \cdot 10^{-4}$	0.831 [0.30]	0.866 [0.42]	0.864 [0.54]

(b) Marginal variance

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.978 [1.6]	0.977 [2.0]	0.976 [2.1]
$5 \cdot 10^{-3}$	0.957 [1.5]	0.974 [1.8]	0.960 [2.2]
$5 \cdot 10^{-4}$	0.949 [1.4]	0.966 [1.7]	0.958 [2.0]

Table S12: Frequentist coverage of the 95% HPD credible intervals for the range and the marginal variance when the true range is $\rho_T = 1$ using PriorPC. The average lengths of the credible intervals are shown in brackets.

(a) Range

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.025	0.927 [7.2]	0.915 [4.8]	0.869 [2.5]	0.690 [1.2]
0.1	0.973 [8.2]	0.961 [5.6]	0.924 [2.7]	0.783 [1.3]
0.4	1.000 [14]	1.000 [8.6]	0.997 [4.0]	0.949 [1.6]
1.6	0.993 [44]	0.994 [22]	0.992 [8.3]	0.990 [2.9]

(b) Marginal variance

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.025	0.946 [6.5]	0.936 [4.3]	0.889 [2.2]	0.666 [0.95]
0.1	0.975 [7.5]	0.980 [4.9]	0.940 [2.4]	0.754 [1.1]
0.4	1.000 [13]	1.000 [7.8]	0.998 [3.4]	0.912 [1.3]
1.6	0.996 [41]	0.986 [20]	0.984 [7.3]	0.999 [2.2]

Table S13: Frequentist coverage of 95% HPD credible intervals for range and marginal variance when the true range $\rho_T = 1$ using PriorUn1, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.980 [1.5]	0.946 [17]	0.979 [179]
$5 \cdot 10^{-3}$	0.989 [1.5]	0.938 [17]	0.967 [178]
$5 \cdot 10^{-4}$	0.979 [1.5]	0.931 [17]	0.967 [178]

(b) Marginal variance

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.977 [1.8]	0.989 [17]	0.996 [176]
$5 \cdot 10^{-3}$	0.979 [1.8]	0.985 [18]	0.991 [175]
$5 \cdot 10^{-4}$	0.979 [1.8]	0.983 [17]	0.987 [177]

Table S14: Frequentist coverage of 95% HPD credible intervals for range and marginal variance when the true range $\rho_T = 1$ using PriorUn2, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$A\backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.945 [1.4]	0.974 [9.8]	0.985 [40]
$5 \cdot 10^{-3}$	0.936 [1.4]	0.959 [9.6]	0.973 [39]
$5 \cdot 10^{-4}$	0.954 [1.4]	0.961 [9.5]	0.966 [39]

(b) Marginal variance

$A\backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.936 [1.6]	0.983 [8.7]	0.987 [35]
$5 \cdot 10^{-3}$	0.933 [1.6]	0.971 [8.6]	0.980 [34]
$5 \cdot 10^{-4}$	0.937 [1.6]	0.969 [8.7]	0.968 [33]

Table S15: Frequentist coverage of the 95% HPD credible intervals for range and marginal variance when the true range is 0.1 and true marginal variance is 1, where the average length of the credible intervals are given in brackets, for the spatial logistic regression example.

(a) Range

$\rho_0\backslash\sigma_0$	40	10	2.5	0.625
0.0025	0.582 [0.22]	0.575 [0.18]	0.577 [0.17]	0.539 [0.15]
0.01	0.922 [0.30]	0.925 [0.28]	0.906 [0.24]	0.883 [0.21]
0.04	0.999 [0.46]	0.999 [0.40]	1.000 [0.33]	0.998 [0.27]
0.16	0.994 [1.0]	0.995 [0.8]	0.983 [0.57]	0.972 [0.38]

(b) Marginal variance

$\rho_0\backslash\sigma_0$	40	10	2.5	0.625
0.0025	0.968 [1.8]	0.945 [1.7]	0.944 [1.6]	0.867 [1.1]
0.01	0.973 [1.9]	0.961 [1.8]	0.954 [1.6]	0.885 [1.1]
0.04	0.982 [2.2]	0.978 [2.1]	0.961 [1.7]	0.893 [1.2]
0.16	0.993 [3.9]	0.991 [3.3]	0.991 [2.3]	0.950 [1.4]

S7.1 Parametrizing the extra flexibility

Instead of adding spatial variation to the coefficients of the SPDE in Equation (5), κ and τ, one can vary the geometry of the space in a similar way as the deformation method (Sampson and Guttorp, 1992). If E is the Euclidean space $\mathbb{R}^2$, the simple SPDE

$$(1 - \Delta_E)u(\boldsymbol{s}) = \sqrt{4\pi}\mathcal{W}_E(\boldsymbol{s}), \quad \boldsymbol{s} \in E, \tag{6}$$

generates a stationary Matérn GRF with range $\rho = \sqrt{8}$, marginal variance $\sigma^2 = 1$, and smoothness $\nu = 1$. We introduce spatially varying distances in the space by giving the space geometric structure according to the metric tensor $\mathbf{g}(\boldsymbol{s}) = R(\boldsymbol{s})^{-2}\mathbf{I}_2$, where $R(\cdot)$ is a strictly positive scalar function. This locally scales distances by a factor $R(\boldsymbol{s})^{-1}$,

$$\mathrm{d}\sigma^2 = \begin{bmatrix}\mathrm{d}s_1 & \mathrm{d}s_2\end{bmatrix} \mathbf{g}(\boldsymbol{s}) \begin{bmatrix}\mathrm{d}s_1 \\ \mathrm{d}s_2\end{bmatrix} = R(\boldsymbol{s})^{-2}(\mathrm{d}s_1^2 + \mathrm{d}s_2^2), \tag{7}$$

where $\mathrm{d}\sigma$ is the line element, and s_1 and s_2 are the two coordinates of $E = \mathbb{R}^2$.

The non-stationarity in the correlation structure is then described through the spatially varying geometry in Equation (7), which results in a curved two-dimensional manifold that must be embedded in a space with dimension higher than 2 to exist in Euclidean space. The resulting spatial field does not have exactly constant marginal variance because the curvature of the space is non-constant unless $R(\cdot)$ does not vary in space, but there will be less interaction between $R(\cdot)$ and the marginal variance than between $\kappa(\cdot)$ and the marginal variance. And when $R(\cdot)$ varies slowly, the variation in marginal variances is small.

We can relate the Laplace-Beltrami operator in E to the usual Laplacian in $\mathbb{R}^2$ through

$$\Delta_E = \frac{1}{\sqrt{\det(g)}}\nabla_{\mathbb{R}^2} \cdot (\sqrt{\det(g)}g^{-1}\nabla_{\mathbb{R}^2}) = R(\boldsymbol{s})^2\Delta_{\mathbb{R}^2},$$

and the Gaussian standard white noise in E to the Gaussian standard white noise in $\mathbb{R}^2$ through

$$\mathcal{W}_E(\boldsymbol{s}) = \det(g)^{1/4}\mathcal{W}_{\mathbb{R}^2}(\boldsymbol{s}) = R(s)^{-1}\mathcal{W}_{\mathbb{R}^2}(\boldsymbol{s}).$$

Thus the equivalent SPDE in $\mathbb{R}^2$ can be written as

$$R(\boldsymbol{s})^{-2}\left[1 - R(\boldsymbol{s})^2\Delta_{\mathbb{R}^2}\right] u(\boldsymbol{s}) = R(s)^{-1}\sqrt{4\pi}\mathcal{W}_{\mathbb{R}^2}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathbb{R}^2,$$

where the first factor is needed because the volume element $\mathrm{d}V_E = \sqrt{\det(g)}\mathrm{d}V_{\mathrm{R}^2}$. The SPDE can be written as

$$(R(s)^{-2} - \Delta_{\mathbb{R}^2})u(\boldsymbol{s}) = \sqrt{4\pi}R(s)^{-1}\mathcal{W}_{\mathbb{R}^2}, \quad \boldsymbol{s} \in \mathbb{R}^2, \tag{8}$$

in Euclidean space, but we can interpret the non-stationarity through the implied metric tensor. The procedure is similar to the simple reparametrization $\kappa(\cdot) = R(\cdot)^{-1}$, but the extra factor on the right-hand side of the equation reduces the variability of the marginal variances due to changes in $\kappa(\cdot)$.

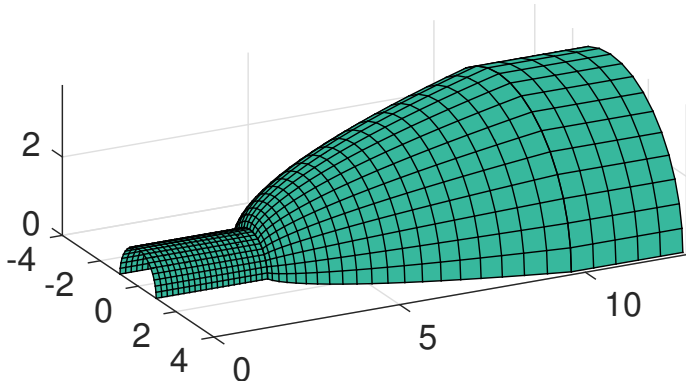


Figure S6: Half cylinder deformed according to the spatially varying metric tensor. The lines formed a regular grid on the half cylinder before deformation.

For example, the space $[0, 9] \times [0, 3]$ with the Euclidean distance metric can be visualized as a rectangle, which exists in $\mathbb{R}^2$, or as a half cylinder with radius $3/\pi$ and height 9, which exists in $\mathbb{R}^3$, but if the space is given the spatially varying metric tensor defined by the local range function

$$R(s_1, s_2) = \begin{cases} 1 & 0 \leq s_1 < 3, 0 \leq s_2 \leq \pi, \\ (s_1 - 2) & 3 \leq s_1 < 6, 0 \leq s_2 \leq \pi, \\ 4 & 6 \leq s_1 \leq 9, 0 \leq s_2 \leq \pi, \end{cases} \tag{9}$$

the space cannot be embedded in $\mathbb{R}^2$. With this metric tensor, the space is no longer flat, and must be embedded in $\mathbb{R}^3$ as, for example, the deformed cylinder shown in Figure S6. Thus, solving Equation (8) with the spatially varying coefficient is the same as solving Equation (6) on the deformed space. This means that unlike the deformation method, a spatially varying $R(\cdot)$ does not correspond to a deformation of $\mathbb{R}^2$ to $\mathbb{R}^2$, but rather from $\mathbb{R}^2$ to a higher-dimensional space.

Since the variation in the marginal variances due to variations in the local ranges is small if $R(\cdot)$ does not vary too much, we introduce a separate function $S(\cdot)$ that controls the marginal standard deviations of the process and limit the SPDE to a region of interest, $\mathcal{D}$, with Neumann boundary conditions,

$$(R(s)^{-2} - \Delta_{\mathbb{R}^2}) \left(\frac{u(\boldsymbol{s})}{S(\boldsymbol{s})} \right) = \sqrt{4\pi} R(s)^{-1} \mathcal{W}_{\mathbb{R}^2}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D}.$$

This introduces boundary effects as was discussed in the paper by Lindgren et al. (2011), but we will not discuss the effects of the boundary in this paper.

This SPDE allows for greater separation of the parameters that affect correlation structure and the parameters that affect marginal standard deviations than the previous approach, and demonstrates the usefulness of careful consideration of how the spatially varying behaviour is introduced and parametrized. The SPDE derived based on the metric tensor allows for separate priors for correlation structure and marginal standard deviations through expansions of $\log(R(\cdot))$ and $\log(S(\cdot))$ into bases.

S7.2 Setting priors on the parameters

There are two sources of non-stationarity in the flexible extension from stationarity: a function $R(\cdot)$ that controls local range and a function $S(\cdot)$ that controls the marginal standard deviation. The degree of flexibility in each of these sources of non-stationarity must be controlled to limit the risk of overfitting. Due to the issues of singular and equivalent Gaussian measures discussed in the main paper, we will not follow the PC prior framework, but instead use a construction motivated by the principles of the PC priors to make the non-stationary model contract towards a base model of stationarity. Denote by $\boldsymbol{\theta}$ the extra parameters added to the GRF that move the model away from the base model of stationarity, $\boldsymbol{\theta} = \mathbf{0}$. The prior on $\boldsymbol{\theta}$ will be constructed conditionally on the parameters of the stationary GRF, ρ and σ^2, and for each choice of these parameters, $\boldsymbol{\theta}$ should shrink towards $\mathbf{0}$.

We parametrize the local distance, $R(\cdot)$, and the approximate marginal standard deviations, $S(\cdot)$, through

$$\begin{aligned} \log(R(\boldsymbol{s})) &= \log\left(\frac{\rho}{\sqrt{8}}\right) + \sum_{i=1}^{n_1} \theta_{1,i} f_{1,i}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D}, \\ \log(S(\boldsymbol{s})) &= \log(\sigma) + \sum_{i=1}^{n_2} \theta_{2,i} f_{2,i}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D}, \end{aligned} \tag{10}$$

where $\{f_{1,i}\}$ is a set of basis functions for the local range centred such that $\langle f_{1,i}, 1\rangle_{\mathcal{D}} = 0$, for $i = 1, 2, \ldots, n_1$, and $\{f_{2,i}\}$ is a set of basis functions for the marginal standard deviations centred such that $\langle f_{2,i}, 1\rangle_{\mathcal{D}} = 0$ for $i = 1, 2, \ldots, n_2$. We collect the parameters in vectors $\boldsymbol{\theta}_1 = (\theta_{1,1}, \ldots, \theta_{1,n_1})$ and $\boldsymbol{\theta}_2 = (\theta_{2,1}, \ldots, \theta_{2,n_2})$ such that $\boldsymbol{\theta}_1$ controls the non-stationarity in the correlation structure and $\boldsymbol{\theta}_2$ controls the non-stationarity in the marginal standard deviations.

We want the prior for each source of non-stationarity to be invariant to scaling of the covariates and to handle linear dependencies between the covariates in a reasonable way, and we follow the basic idea of the g-priors (Zellner, 1986) (with $g = 1$),

$$\boldsymbol{\theta}_1|\tau_1 \sim \mathcal{N}(\mathbf{0}, \tau_1^{-1}\mathbf{S}_1^{-1}) \quad \text{and} \quad \boldsymbol{\theta}_2|\tau_2 \sim \mathcal{N}(\mathbf{0}, \tau_2^{-1}\mathbf{S}_2^{-1}),$$

where S_1 is the Gramian,

$$S_{1,i,j} = \frac{\langle f_{1,i}, f_{1,j}\rangle_{\mathcal{D}}}{\langle 1, 1\rangle_{\mathcal{D}}}, \quad \text{for } i, j = 1, 2, \ldots, n_1,$$

and S_2 is similarly the Gramian based on $\{f_{2,i}\}$. In this set-up the Gramians account for the structure in the basis functions and the strictness of the priors are controlled by two precisions parameters τ_1 and τ_2. If the precision parameters are fixed hyperparameters, the resulting priors are Gaussian. However, the Gaussian probability density is flat at zero due to the infinite differentiability of the density function, and we prefer a prior that has a spike at zero.

This can be achieved by selecting the hyperpriors to be the PC prior for the precision parameter in a Gaussian distribution (Simpson et al., 2017), which is designed to shrink towards a base model of zero variance. We combine the selection for the hyperpriors with an *a priori* ansatz that the independence between the correlation structure and the marginal variance in the prior for the stationary model also can be applied to the non-stationarity,

$$\pi(\tau_1) = \frac{\lambda_1}{2}\tau_1^{-3/2}\exp\left(-\lambda\tau_1^{-1/2}\right) \text{ and } \pi(\tau_2) = \frac{\lambda_2}{2}\tau_2^{-3/2}\exp\left(-\lambda_2\tau_2^{-1/2}\right).$$

These hyperpriors for the precision parameters have so heavy tails that integrating them out will introduce infinite spikes in the marginal priors for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ at zero.

The hyperparameters λ_1 and λ_2 control the spread of the priors and can be selected either based on expert knowledge or on frequentist properties. The parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ give multiplicative effects to local range and marginal standard deviations, respectively, and one possibility is to control the size of the multiplicative effect through

$$\text{Prob}\left(\max_{\boldsymbol{s}\in\mathcal{D}}\left|\log\left(\frac{R(\boldsymbol{s})}{\rho/\sqrt{8}}\right)\right| > C_1\middle|\rho,\sigma^2\right) = \text{Prob}\left(\max_{\boldsymbol{s}\in\mathcal{D}}\left|\log\left(\frac{R(\boldsymbol{s})}{\rho/\sqrt{8}}\right)\right| > C_1\right) = \alpha_1,$$

$$\text{Prob}\left(\max_{\boldsymbol{s}\in\mathcal{D}}\left|\log\left(\frac{S(\boldsymbol{s})}{\sigma^2}\right)\right| > C_2\middle|\rho,\sigma^2\right) = \text{Prob}\left(\max_{\boldsymbol{s}\in\mathcal{D}}\left|\log\left(\frac{S(\boldsymbol{s})}{\sigma^2}\right)\right| > C_2\right) = \alpha_2.$$

One can see from Equation (10) that the relative differences do not depend on the parameters of the stationary model, and the full prior factors as $\pi(\rho,\sigma^2,\boldsymbol{\theta}) = \pi(\rho)\pi(\sigma^2)\pi(\boldsymbol{\theta}_1)\pi(\boldsymbol{\theta}_2)$.

In practice, it is difficult to have an informed, *a priori* opinion on the non-stationary part of the model, but the hyperparameters λ_1 and λ_2 can be chosen in such a way that they give a conservative prior. Since stationarity is our base model and the non-stationarity is provided as extra flexibility, we will require that the hyperparameters are set such that the inference behaves well when the true data-generating distribution is stationary. We propose to set the hyperparameter by first fitting the stationary model, using the maximum aposteriori estimate of the parameters to make multiple simulated datasets from the stationary GRF and nugget effect, fit the non-stationary GRF with a nugget effect to each dataset, and calculate the frequentist coverage of the non-stationarity parameters. The hyperparameters can then be set such that the coverage of the credible intervals of the non-stationary parameters is close to nominal coverage. This ensures that the prior provides enough regularization that each posterior marginal for the non-stationarity parameters do not suggest non-stationarity when a stationary data-generating function is used.

S7.3 Supplementary details for example on precipitation

The example is described in the main paper and this section provides extra figures and details that supplements the presentation in the main paper.

In the SPDE approach the spatial field is defined on a triangular mesh and the values of the spatial field within the triangles are defined through linear interpolation based on

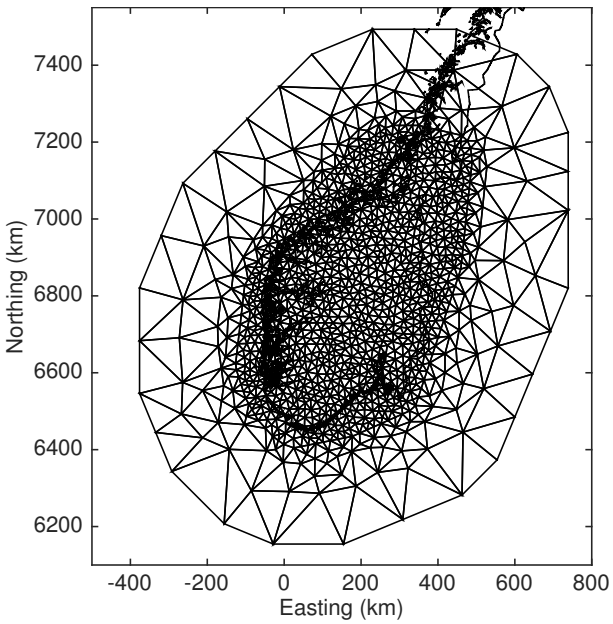


Figure S7: Mesh used for the SPDE approach.

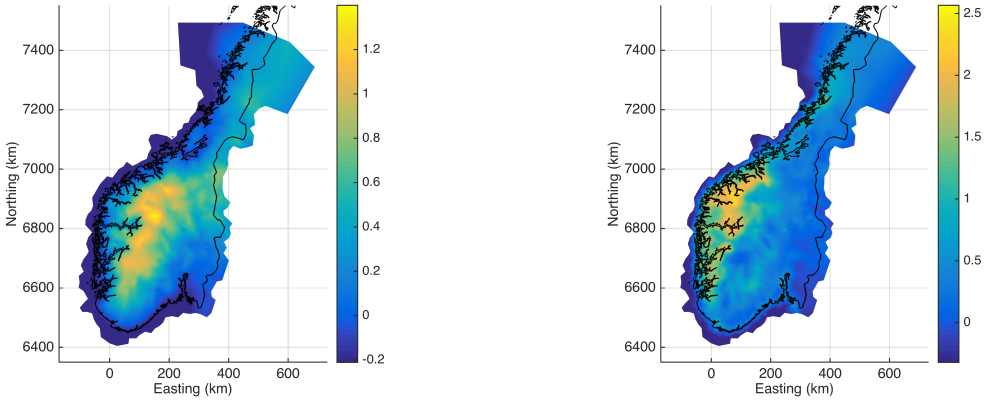
the values on the nodes of the mesh. This means that the elevation covariate in the first-order structure is only needed at each observation location, but that the elevation and gradient covariates in the second-order structure are needed at every location within the triangulation. We use the mesh shown in Figure S7 and project elevation and gradient values from the high resolution digital elevation map GLOBE (Hastings et al., 1999) onto the mesh. The projection is piece-wice linear on each triangle of the mesh and minimizes the integrated square deviation over the domain covered by the mesh. This results in the piece-wise linear covariates shown in Figure S8.

The coefficients, $\boldsymbol{\theta}_1$, of the two linear covariates in $\log(R(\cdot))$ are given the prior

$$\begin{aligned}\boldsymbol{\theta}_1|\tau_1 &\sim \mathcal{N}(\mathbf{0}, S_1/\sqrt{\tau_1})\\ \tau_1 &\sim \frac{\lambda_1}{2}\tau_1^{-3/2}\mathrm{e}^{-\lambda_1/\sqrt{\tau_1}}\end{aligned}$$

as described in the previous section, and the coefficients, $\boldsymbol{\theta}_2$, of the two linear covariates in $\log(S(\cdot))$ are given a similar prior, but with hyperparameter λ_2. The non-stationary model is more difficult to fit in the INLA framework than the stationary model because the priors for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ have infinite spikes in $\mathbf{0}$ that makes the posteriors non-Gaussian in the area around the origin. The optimization can be improved by reparametrizing as $\boldsymbol{\theta}'_1 = \boldsymbol{\theta}_1\sqrt{\tau}_1$ and $\boldsymbol{\theta}'_2 = \boldsymbol{\theta}_2\sqrt{\tau}_2$, but the marginal posteriors will not be sufficiently peaked at the origin and will miss the multimodality that should be present when there is a mode close to zero. However, we still use INLA as a fast approximation for the repeated fitting of the datasets needed for selecting the hyperparameters of the prior for non-stationarity.

The non-stationary model was fitted using an MCMC sampler and the resulting



(a) Elevation (km)

(b) Magnitude of gradient (100m/km)

Figure S8: The covariates (a) elevation and (b) magnitude of the gradient used for the covariance structure.

posterior means of the range and the standard deviation are shown in Figure S9. From Figure S10 one can see that the spatially varying range and standard deviation leads to non-stationarity in the correlation structure and the marginal standard deviations of the spatial effect. However, the effect in standard deviations appear to be stronger than the effect of the spatially varying range. The posteriors for the multiplicative effects on the stationary range and standard deviation for the western location in Figure S10a shown in Figure S11 shows that the effects are significant in that location. The posterior probabilities for the effects to be less than 1 and greater than 1 are 99% and 99%, respectively. This shows that the the more flexible non-stationary model is preferring to move away from the stationary model even under a conservatively selected prior.

S8 Additional theorem

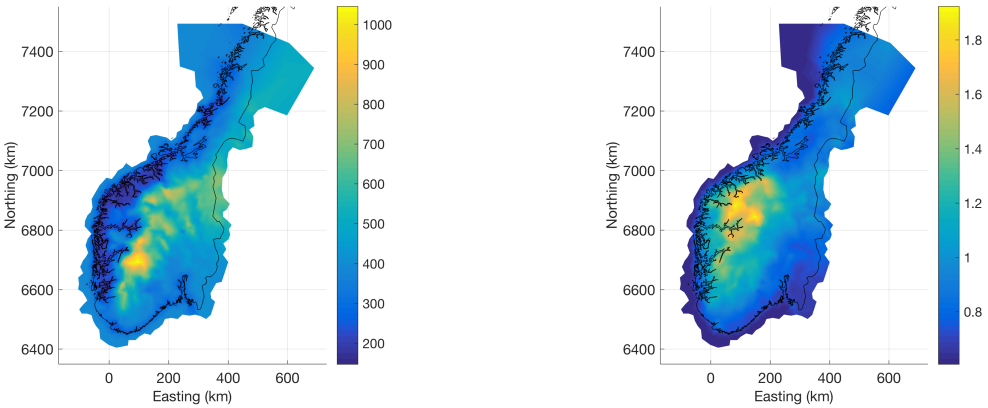
In the proof of the main result in the paper it is necessary to show that the integral used to calculate the KLD is finite. The following theorem shows that this holds in dimensions $d = 1$, $d = 2$ and $d = 3$.

Theorem S8.1. *The definite integral*

$$I_d = \int_{\mathbb{R}^d} \left[\left(\frac{||\boldsymbol{w}||^2}{(1 + ||\boldsymbol{w}||^2)} \right)^{\alpha} - 1 - \log \left(\frac{||\boldsymbol{w}||^2}{(1 + ||\boldsymbol{w}||^2)} \right)^{\alpha} \right] \mathrm{d}\boldsymbol{w},$$

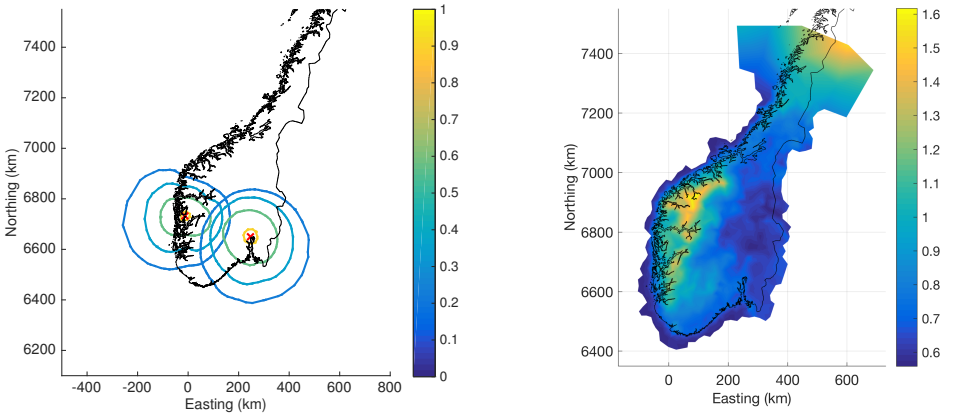
where $\alpha > 0$, is finite for $d \leq 3$.

Proof. The definite integral can be expressed an an integral in d-dimensional spherical



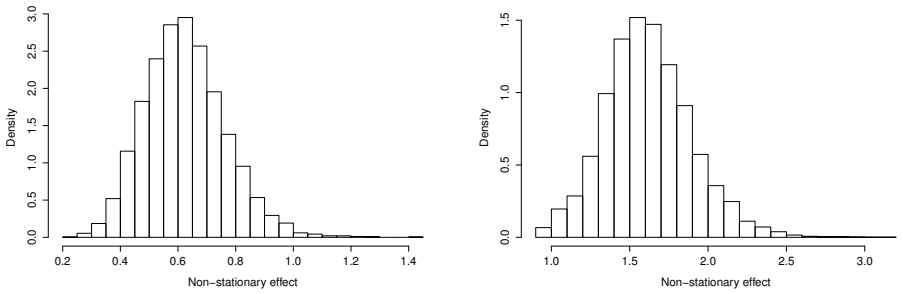
(a) Range (b) Standard deviation

Figure S9: Posterior mean of (a) range and (b) standard deviation.



(a) Level curves of correlations (b) Marginal standard deviations

Figure S10: Covariance structure described through (a) 0.90, 0.57, 0.36 and 0.22 level curves of correlation with respect to the two locations marked with red crosses and (b) marginal standard deviations.



(a) Multiplicative effect on range

(b) Multiplicative effect on standard deviation

Figure S11: Posteriors of the multiplicative effect on the stationary (a) range and (b) standard deviation at the western location in Figure S10a.

coordinates,

$$I_d = C_d \int_0^\infty \left[\left(\frac{r^2}{1+r^2} \right)^\alpha - 1 - \log \left(\frac{r^2}{1+r^2} \right)^\alpha \right] r^{d-1} \mathrm{d}r, \tag{11}$$

where C_d is a finite constant that varies with dimension. There are two issues: the behaviour for small r and the behaviour for large r. For $d = 1$,

$$0 \leq I_d \leq -C_1 \alpha \int_0^\infty \log \frac{r^2}{1+r^2} \mathrm{d}r = \pi \alpha C_1 < \infty,$$

and the definite integral is finite for $d = 1$. Furthermore, the factor r^{d-1} makes the value of the integrand smaller close to 0 for larger d and we can conclude that the behaviour around 0 is not a problem for any $d \geq 1$.

The behaviour of the integrand for large r can be studied through an expansion of the integrand in $(1+r^2)^{-1}$. The part between the square brackets in Equation (11) behaves as

$$\frac{\alpha^2}{2} \frac{1}{(1+r^2)^2} + \mathcal{O}\left(\frac{1}{(1+r^2)^3} \right).$$

This means that there exists a $0 < r_0 < \infty$ such that

$$\begin{aligned} &\int_0^\infty \left[\left(\frac{r^2}{1+r^2} \right)^\alpha - 1 - \log \left(\frac{r^2}{1+r^2} \right)^\alpha \right] r^{d-1} \mathrm{d}r \\ &\qquad \leq C_1 + \int_{r_0}^\infty \left[\frac{\alpha^2}{2} \frac{1}{(1+r^2)^2} + \frac{C_2}{(1+r^2)^3} \right] r^{d-1} \mathrm{d}r, \end{aligned}$$

where $|C_1| \leq \infty$ due to the finiteness for $d = 1$, and C_2 is a constant. For $d \leq 3$ both terms on the right hand side are finite. Thus I_d is finite for $d \leq 3$. □

References

Berger, J. O., De Oliveira, V., and Sansó, B. (2001). Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96(456):1361–1374.

Bogachev, V. I. (1998). *Gaussian measures*. Number 62 in Mathematical Surveys and Monographs. American Mathematical Soc.

Chen, M.-H. and Shao, Q.-M. (1999). Monte carlo estimation of bayesian credible and hpd intervals. *Journal of Computational and Graphical Statistics*, 8(1):69–92.

Hastings, D. A., Dunbar, P. K., Elphingstone, G. M., Bootz, M., Murakami, H., Maruyama, H., Masaharu, H., Holland, P., Payne, J., Bryant, N. A., Logan, T. L., Muller, J.-P., Schreier, G., and MacDonald, J. S. (1999). The global land one-kilometer base elevation (globe) digital elevation model, version 1.0.

Ingebrigtsen, R., Lindgren, F., and Steinsland, I. (2014). Spatial models with explanatory variables in the dependence structure. *Spatial Statistics*, 8:20–38.

Ingebrigtsen, R., Lindgren, F., Steinsland, I., and Martino, S. (2015). Estimation of a non-stationary model for annual precipitation in southern norway using replicates of the spatial field. *Spatial Statistics*, 14, Part C:338–364.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Sampson, P. D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.*, 32(1):1–28.

Smith, B. J. et al. (2007). boa: an r package for mcmc output convergence assessment and posterior inference. *Journal of Statistical Software*, 21(11):1–37.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243.