ISBN 978-82-326-0920-8 (printed version)
ISBN 978-82-326-0921-5 (electronic version)
ISSN 1503-8181

Doctoral theses at NTNU, 2015:132

Geir-Arne Fuglstad

Modelling Spatial Non-stationarity

NTNU
Norwegian University of
Science and Technology
Faculty of Information Technology,
Mathematics and Electrical Engineering
Department of Mathematical Sciences

NTNU – Trondheim
Norwegian University of
Science and Technology

Doctoral theses at NTNU, 2015:132

NTNU

NTNU – Trondheim
Norwegian University of
Science and Technology

Geir-Arne Fuglstad

Modelling Spatial Non-stationarity

Thesis for the degree of Philosophiae Doctor

Trondheim, June 2015

Norwegian University of Science and Technology

NTNU – Trondheim
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

ISBN 978-82-326-0920-8 (printed version)
ISBN 978-82-326-0921-5 (electronic version)
ISSN 1503-8181

Doctoral theses at NTNU, 2015:132

Printed by Skipnes Kommunikasjon as

Preface

This thesis was submitted for partial fulfillment of the requirements for the Ph.D. degree at the Department of Mathematical Sciences at the Norwegian University of Science and Technology (NTNU). The research was completed at NTNU during the years 2011–2015.

I would like to thank my supervisors Professor Håvard Rue at NTNU, Dr. Finn Lindgren at the University of Bath, and Dr. Daniel Simpson at the University of Warwick. They have offered invaluable experience and advice for the work and the completion of the thesis. I would also like to thank my friends and colleagues for making it a wonderful experience.

Trondheim
March 2015

Geir-Arne Fuglstad

1 Introduction

Throughout the entire field of spatial statistics there is always an underlying feeling that stationary models, which have the same spatial dependence structure everywhere, cannot possibly be true and that it should be possible to do better by accounting for the lack of stationarity. This thought provides the main motivation for the work on non-stationary models, but, in practice, the dichotomy between stationary models and non-stationary models is not as rigid as it might seem. There are unimaginably many ways to introduce non-stationarity, and the question of which type of non-stationarity that should be included in the model becomes equally important as the question of whether one should use a non-stationary model or not.

It is more difficult to estimate the dependence structure of a spatial model than the mean structure, and there will be a tendency for the estimated dependence structure to depend on the model used. Even if we assume that we have a single realization where the mean of the spatial process is known, the covariances estimated between the observed locations are dependent on the covariance model chosen for the estimation. And when there are multiple realizations, empirical covariances are available between the observed locations, but the model must still fill the missing covariances between the unobserved locations. The difficulty seems intuitively reasonable from a hierarchical modelling viewpoint since the covariance structure is below the observation level and the spatial process level in the hierarchical model, and requires information to propagate several levels from the observed data.

The overall goal of this thesis is to contribute new, useful, computationally efficient models that can account for spatially varying dependence structure in spatial problems. But it is difficult to extract information about a spatially varying dependence structure and the dependence structure has an inherent lack of identifiability. Therefore, we would like to fit the models within a Bayesian framework to stabilise the inference, understand better the *a priori* assumptions we put into the model and account properly for uncertainty, but we do not yet know how to set sensible priors on non-stationarity. This has lead to a division of the thesis into to two strongly connected parts: the development of the models and the development of sensible priors.

The priors and the hyperparameters are important for non-stationary models. This has been noted, for example, by Neto et al. (2014) who observed that the non-stationary model of Paciorek and Schervish (2006) was sensitive to the choice of hyperparameters. It is possible to improve the situation by reducing the dimensionality of the non-stationarity by only using a couple of covariates, but even for these models, prior sensitivity and computational problems have been reported (Ingebrigtsen et al., 2014a,b). But before attacking this problem, we need to tackle one of the fundamental problems in Bayesian spatial statistics that continues to trouble statisticians to this day: which prior should we use for the range? The answer to this question should form the foundation for the work on priors for non-stationarity.

In the next sections we introduce the main concepts of spatial statistics, and discuss the main challenges of non-stationary modelling and the contributions made by the papers in this thesis. Then in Paper I we introduce a new non-stationary model and explore the possibilities for representing the anisotropy in a non-stationary model through a vector field. In Paper II we present a large case study where we apply this model to annual precipitation in the conterminous US and discuss the issue of how the non-stationarity is best captured by the proposed model. The thesis ends with Paper III, in which we discuss a new simple and theoretically justified joint prior for the range and the marginal variance in spatial models, and extensions to priors for non-stationary models. Together the papers in this thesis provide a solid foundation for developing sensible priors for non-stationarity and extending the proposed models to a Bayesian framework.

1.1 Spatial statistics

One of the simplest statistical models is the simple linear regression model where each observation y_i, for $i = 1, 2, \ldots, N$, has an associated covariate x_i and is assumed to follow the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where β_0 is the intercept, β_1 is the slope, and each ϵ_i is independently identically distributed as $\mathcal{N}(0, \sigma^2)$. This model assumes that when the mean

structure is known, the errors are independent, but this is an assumption that is rarely true for spatial data. For spatial data the observations tend to exhibit dependence simply because they are close to each other. This might, in some cases, be caused by spatially varying structure in the mean that is not captured in the model, but even after we have included all the covariates we believe are relevant, there tends to be spatial dependence in the residuals.

This motivates the inclusion of another error term in the regression equation,

$$y_i = \beta_0 + \beta_1 x_i + u(\boldsymbol{s}_i) + \epsilon_i, \tag{1}$$

denoted by $u(\cdot)$, which is a spatially correlated random error where close locations are more correlated than distant locations. In this model there is a spatial dependence between the residuals of the observations even after the mean structure is known. The inclusion of spatially correlated effects is what characterizes the field of spatial statistics and requires the construction of random objects that can generate spatially varying random fields.

The most common such random object for continuously indexed phenomena is called a Gaussian random field and is an extension of the finite-dimensional multivariate Gaussian distributions to temporal and spatial random fields.

Definition 1.1 (Gaussian random field (GRF))**.** Let $\mathcal{D} \subset \mathbb{R}^n$, then the random field $\{u(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ is said to be a Gaussian random field if for all $m \in \mathbb{N}$ for all choices of points $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m \in \mathcal{D}$, $(u(\boldsymbol{s}_1), \ldots, u(\boldsymbol{s}_m))$ has a multivariate Gaussian distribution.

We will ignore the technicalities associated with sample path properties and consider a GRF to be uniquely defined by the mean value at each location and the covariance between each pair of locations. These properties are usually described by a mean value function and a covariance function.

Definition 1.2 (Covariance and mean value functions)**.** Let $\{u(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ be a GRF, then the function $c : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ defined by

$$c(\boldsymbol{s}_1, \boldsymbol{s}_2) = \mathrm{Cov}[u(\boldsymbol{s}_1), u(\boldsymbol{s}_2)]$$

is called the covariance function of the process and the function $m : \mathcal{D} \to \mathbb{R}$ defined by

$$m(\boldsymbol{s}) = \mathrm{E}[u(\boldsymbol{s})]$$

is called the mean value function of the process.

In classical geostatistical models the covariance function is assumed to be *isotropic*, which means that the covariances only depend on the distances between the locations, $c(\boldsymbol{s}_1, \boldsymbol{s}_2) = r(||\boldsymbol{s}_2 - \boldsymbol{s}_1||)$. The most commonly used family of isotropic covariance functions is the Matérn family (Matérn et al., 1960),

$$r(d) = \sigma^2 \frac{1}{2^{\nu-1}\Gamma(\nu)} (\kappa d)^\nu K_\nu(\kappa d),$$

where κ is the inverse range parameter, σ^2 is the marginal variance, ν is the smoothness, and K_ν is the modified Bessel function of second kind, order ν. This family of covariance functions is considered to be a general class of covariance functions that possess useful and good properties (Handcock and Stein, 1993) and Stein (1999) goes as far as to strongly advocate the use of this family of covariance functions.

GRFs are common building blocks in spatial models and when included in generalized linear models they offer a flexible modelling tool for many different types of data. The set-up in Equation (1) is aimed at continuously indexed data, but there also exist other types of data such as discretely indexed data and spatial point process data. The continuously indexed models and the discretely indexed models have, traditionally, developed separately with the continuously indexed models having appealing theoretical properties and the discretely indexed models having appealing computational properties.

The mean value function and the covariance function describe a continuously indexed random object, but in computations we only observe a finite number of points. In some cases, discretely indexed data can be handled by *Gaussian Markov random fields* (GMRFs) (Rue and Held, 2005), which are multivariate Gaussian distributions with sparse precision matrices that can be exploited for fast computations. GMRFs have good computational properties, but they are hard to specify in a spatially consistent way for irregularly spaced data and thus not directly applicable for these type of data.

However, Lindgren et al. (2011) showed that for the Matérn models it is possible to make a connection between the continuously indexed model and discretely indexed GMRFs through the use of a stochastic partial differential equation (SPDE). This connection makes it possible to combine the best of two worlds: good theoretical properties and good computational properties.

1.2 Computationally efficient spatial statistics

The Matérn covariance function describes the entire second-order structure of a GRF, but if we assume the classical geostatistical model in Equation (1) and use the direct approach for computations, we are forced to form a dense $N \times N$ covariance matrix Σ containing all the pairwise covariances

$$\Sigma_{i,j} = r(||\boldsymbol{s}_i - \boldsymbol{s}_j||),$$

where $\boldsymbol{s}_i$, for $i = 1, 2, \ldots, N$, are the spatial locations of the observations. The likelihood computations require the evaluation of the determinant of the covariance matrix and the solution of linear systems with the covariance matrix, and lead to a computational complexity $\mathcal{O}(N^3)$ and is the cause of the so-called "big-N problem" in spatial statistics. This has lead to several approaches for making spatial models computationally feasible (Cressie and Johannesson, 2008; Banerjee et al., 2008; Furrer et al., 2006; Stein et al., 2004; Fuentes, 2007; Sun et al., 2012), but the starting point of the work in this thesis is the SPDE approach introduced by Lindgren et al. (2011), which at its core is an efficient way of doing computations with a Matérn GRF, $u(\cdot)$.

The SPDE approach makes the unorthodox decision to throw away the explicit covariance function and instead describe the covariance structure through an SPDE. It was discovered already by Whittle (1954, 1963) that a Matérn GRF can be generated by the SPDE

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

where $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ is the Laplacian, $\alpha = \nu + 1$, and $\mathcal{W}$ is standard Gaussian white noise. The parametrization through κ, τ and α differs from the standard parametrization through κ, σ^2 and ν, but there is a one-to-one correspondence between the two parametrizations.

The novelty in Lindgren et al. (2011) is the combination of the above SPDE with a finite element method (FEM) for solving it. The domain $\mathcal{D} \subset \mathbb{R}^2$ of interest is first triangulated and then the GRF is expanded into a piecewise-linear basis through

$$u(\boldsymbol{s}) = \sum_{k=1}^{M} w_k u_k(\boldsymbol{s}),$$

where the weights $\boldsymbol{w} = (w_1, w_2, \ldots, w_M)$ have a multivariate Gaussian distribution, and the basis functions $\{u_1(\cdot), \ldots, u_M(\cdot)\}$ are constructed such that for each node i in the triangulation the basis function $u_i(\cdot)$ takes the value 1 on node i, the value 0 on all other nodes and is linear on each triangle in the triangulation. This leads to a compactly supported basis where two basis functions only can be non-zero at the same time if they correspond to nodes in the same triangle.

The computational benefits of this approach arise from the selection of the precision matrix $\mathbf{Q}$ in

$$\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}^{-1}).$$

The FEM leads to a spatial sparsity structure in $\mathbf{Q}$ that is essential for the computational efficiency of the method. An element $Q_{i,j}$ is non-zero only for neighbours of order up to and including α. This structure can be exploited through a Cholesky decomposition of $\mathbf{Q}$ to reduce the computational complexity of the calculations from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^{1.5})$. The spatial models are implemented in the R package `INLA` (Rue et al., 2009) for Bayesian inference.

The SPDE approach can be viewed purely as a computational tool that makes spatial models feasible for realistically sized datasets, but this is an unnecessarily limited viewpoint since it provides a useful tool for solving difficult problems such as constructing GRFs on spheres with natural distance measures, multivariate GRFs (Hu et al., 2013b), GRFs with oscillating covariance functions (Hu et al., 2013a), and non-stationary GRFs, which is the main focus of this thesis.

The SPDE in Equation (2) describes an isotropic process, but this can be relaxed to geometric anisotropy by replacing the Laplacian operator with a non-isotropic operator. This can be done by re-writing Equation (2)

as

$$(\kappa^2 - \nabla \cdot \mathbf{H}\nabla)^{\alpha/2} u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathbb{R}^2,$$

where $\mathbf{H}$ is a positive definite matrix. This matrix gives an anisotropic distance measure

$$\text{distance}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \sqrt{(\boldsymbol{s}_2 - \boldsymbol{s}_1)^{\mathrm{T}} \mathbf{H}^{-1} (\boldsymbol{s}_2 - \boldsymbol{s}_1)}$$

and provides a Matérn covariance function with geometric anisotropy. This can be taken one step further by allowing κ and $\mathbf{H}$ to vary spatially. This idea is explored in Papers I and II.

1.3 Background on non-stationarity

The assumption of isotropy in spatial models is in most cases an assumption made out of convenience. There are infinitely many ways a covariance structure could deviate from isotropy, and one could imagine stationarity like a single point in a continuum of models. But the extension to non-stationary spatial fields is non-trivial, and requires that one decides what type of non-stationarity one wants to introduce in the model. Does one want non-stationarity in the marginal variances, non-stationarity in the correlation structure, non-stationarity in the nugget effect, non-stationarity in the smoothness, or maybe some combination of these types of non-stationarities? Further, should the non-stationarity vary smoothly over the region or should there be discontinuities in the non-stationarity? And how does one introduce the non-stationarity in the covariance structure in practice?

Even if all the previous questions could be answered satisfactorily, there is a fundamental inseparability between non-stationarity in the mean and non-stationarity in the covariance structure in geostatistical models when only a single realization is available, which often is the case with spatial data. In classical geostatistical models the non-stationarity in the mean is handled through linear covariates, but the non-stationarity in the covariance structure is ignored. For Gaussian models this does not tend to introduce much error in the point predictions, but the prediction errors are affected and the predictive distributions will be wrong. In the early literature on non-stationary methods the data typically had low spatial dimension and high temporal dimension. This made the separability of

non-stationarity in the mean structure and the covariance structure less of a problem, but the amount of information available on spatial structure from a low spatial dimension is limited. Strictly speaking, the data can only inform about the covariances between observed locations and the rest of the covariances are filled according to the assumptions we put into the model. As a result, the estimated covariance structure will be greatly affected by the choice of model.

The well-known deformation method of Sampson and Guttorp (1992) fills the unobserved structure by assuming that the underlying spatial process can be made stationary through a spatial deformation of the domain. The assumption of the existence of a spatial deformation by itself is not enough, and it must be combined with a parametric form for the spatial deformation and a penalty parameter controlling how much the deformation can vary from the identity deformation. This is a flexible model where the estimated spatial structure can be visualized from a visualization of the spatial deformation, but it is has several issues. There is a possibility of folding whereby two locations are mapped to the same location and it is computationally heavy to estimate spatial deformations even for a low number of spatial locations. The original formulation of the deformation method was entirely based on the empirical covariances and used no distributional assumption for the observations, but the method was later extended to a Bayesian framework by Damian et al. (2001, 2003) and Schmidt and O'Hagan (2003). These models all require replicated observations and it was not until several years later that the methodology was extended to a single realization by Anderes and Stein (2008). Recently, there have been developments into covariates in the covariance structure (Schmidt et al., 2011), and to higher dimensional base spaces (Bornn et al., 2012).

The other well-known, classical method for constructing non-stationary spatial fields constructs a non-stationary spatial field by convolving a spatially varying kernel with white noise. This approach is commonly called the process convolution approach and started with the works of Higdon (1998) and Higdon et al. (1999), but it was later discovered that there exists a specific choice of the kernel function which leads to a closed form covariance function (Paciorek and Schervish, 2006). The covariance function is a Matérn-like covariance function where the distance measure is

defined through the arithmetic mean of positive definite matrices,

$$\text{distance}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \sqrt{(\boldsymbol{s}_2 - \boldsymbol{s}_1)^{\mathrm{T}} \left(\frac{\Sigma(\boldsymbol{s}_1) + \Sigma(\boldsymbol{s}_2)}{2} \right)^{-1} (\boldsymbol{s}_2 - \boldsymbol{s}_1)},$$

where $\boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathbb{R}^2$ are two arbitrary spatial locations, and $\Sigma(\boldsymbol{s}_1)$ and $\Sigma(\boldsymbol{s}_2)$ are positive definite matrices at locations $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, respectively. The availability of the covariance function in closed form can be considered a positive feature of the model since covariance functions are a familiar concept, but the properties of the average of positive definite matrices can be non-intuitive. If the domain consists of two parts where the left side has long dependence in the vertical direction and the right side has long dependence in the horizontal direction, the distance between a location in the left side and a location in the right side behaves according to an isotropic distance. This contradicts the intuition that the distance should combine long vertical range on the left side with long horizontal range of the right side in calculating the distance. The model does not include any way to overcome the inherent computational problems of spatial methods and is limited to a low number of spatial locations. The model described by Paciorek and Schervish (2006) uses latent spatial fields to control the spatially varying covariance structure, but it is also possible to use covariates to control the spatially varying covariance structure (Neto et al., 2014).

1.4 Challenges in non-stationary spatial modelling

The field of non-stationary spatial modelling has several unsolved challenges such as the increased computational effort required compared with stationary models, how to compare different methods and their estimated covariance structures, how to visualize an estimated covariance structure, how to visualize the uncertainty of the estimated covariance structure, and how to set sensible priors on the non-stationarity. To provide solutions to all of these challenges would go far beyond any single thesis and the papers in this thesis are focused on the computational aspect and on the issue of priors.

There does not tend to be any comparison of the results of new methods with the results from old methods because the methods for non-stationarity

are difficult to implement and there is a lack of availability of code, and because the methods require tuning and long computation times. Further, it is not clear how to compare the results from two different models beyond predictive measures. It is difficult to visualize a non-stationary covariance structure and there are no clear ways to compare covariance functions or to visualize the difference between them. Making comparisons with other methods possible requires effort from the creators of the different methods and there was an initiative for improving the situation with a workshop on non-stationary covariance functions at PASI 2014[1], but this remains a huge challenge. Additionally, the standard deformation method and process convolution method cannot handle the number of observations used in Paper II and we were forced, as most other authors, to compare with stationary models and different settings for the proposed method.

One way to overcome the computational issues of non-stationary methods is to combine an approach for non-stationarity with techniques like covariance tapering to improve the computation times (Katzfuss, 2013), but much of the literature has been focused on constructing complex, flexible models for non-stationarity and not on increasing the computational feasibility of the methods. The size of the problems that can be solved has improved due to the improvements in computation hardware, but the sizes of the available datasets have similarly increased and waiting for better computers is not a satisfactory solution. Recently, there has been a move away from general flexible models toward models where the covariance structure is controlled by a small number of covariates (Schmidt et al., 2011; Ingebrigtsen et al., 2014a,b; Neto et al., 2014), but reducing the number of parameters is not enough to solve the inherent problem of spatial statistics, which is that the calculations scale cubically in the number of observation locations.

The question of priors for non-stationarity has received little attention, and in fact even for stationary models there is no consensus on which priors to use. The common approach in non-stationary models controlled by covariates is to use simple priors such as Gaussian distributions and in flexible non-stationary models it is common to use transformed Gaussian process priors on the underlying spatially varying parameters. However, these Gaussian process priors can be sensitive to the hyperpriors and re-

[1]Pan-American Advanced Study Institute on Spatio-Temporal Statistics

quires careful selection of the hyperparameters for convergence.

Perhaps the most important question in non-stationary modelling from an applied viewpoint is when are the non-stationary models worth the extra effort? If there is only a marginal improvement in the predictive distributions, it is hard to justify the extra time needed both to understand and implement the method, and to run the models. In some cases, it is possible to perform tests for non-stationarity (Bowman and Crujeiras, 2013; Jun and Genton, 2012; Fuentes, 2005), but even if evidence of non-stationarity is present, it is interesting to know how much this affects the predictive distributions.

1.5 Contributions of the thesis

The thesis is based on the following papers.

Paper I: Fuglstad, G.-A., Lindgren, F., Simpson, D., and Rue, H. (2015). Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy. *Statistica Sinica*, 25:115–133.

Paper II: Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2015). Does non-stationary spatial data always require non-stationary random fields?. *In revision.*

Paper III: Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2015). Interpretable Priors for Hyperparameters for Gaussian Random Fields. *Technical Report.*

The starting point of Paper I was to extend the recent work by Lindgren et al. (2011), which describes a computationally efficient method for Gaussian Matérn fields, to non-stationary GRFs. The main goals were to preserve the good computational properties of the SPDE approach in a non-stationary GRF and to exploit the local definition of SPDEs to do the non-stationary modelling in a local way by varying the coefficients of the SPDE over the domain. The main contributions of the paper are to describe in detail the application of the finite volume method to construct a computationally efficient discretization of the SPDE and a proof-of-concept of the approach of controlling the range and anisotropy of the spatial field through a vector field. The method presented in Paper I still missed some

key features such as control of the marginal variances together with the anisotropy and a demonstration of the inference scheme on a real-world example.

Thus in Paper II we extended the method in Paper I with the missing control of marginal variances and performed a large case study of annual precipitation in the conterminous US with focus on how to model the apparent non-stationarity in the data. The paper shows that blind application of general flexible models for non-stationarity runs the risk of estimating a covariance structure that to a high degree differs from stationarity due to uncaptured behaviour in the mean structure or types of non-stationarity in the covariance structure that the model cannot capture in a good way. We find that the use of a simple model where the only non-stationarity is a different nugget effect for the western and the eastern regions gives predictions that are of similar quality as a computationally much heavier general flexible model for non-stationarity. This demonstrates the need for methods to determine what type of non-stationarity the data exhibit.

One feature of the model discussed in Papers I and II that is shared with the other approaches for non-stationarity using the SPDE approach (Bolin and Lindgren, 2011; Ingebrigtsen et al., 2014a,b) is the lack of a covariance function in closed form. In these models the implied covariance structure must be understood through the SPDE which describes the model. This way of modelling is unfamiliar to many and there is a need to educate users in understanding models in different ways than through covariance functions. But as discussed by Simpson et al. (2012), the move away from covariance functions is a useful tool for avoiding the computational problems of spatial statistics.

Further, the methods in Papers I and II could also be used with covariates in a similar way as Ingebrigtsen et al. (2014a,b), but our approach has a different aim than the covariate approaches. To use covariates the user has to decide which covariates are important and in which way the covariates moves the model away from stationarity. The model described in the two first papers instead allow the underlying parameters to vary spatially under a spline-like penalty that ensures that the functions behave regularly. This approach is useful when it is not known which covariates should affect the non-stationarity, and the estimated structure could inform about

potential covariates that can be used in the covariance structure. However, if covariates are available, it will always be more computationally efficient to use a covariate-based model, which has far less parameters.

The inference in Papers I and II is done either through empirical Bayes or through penalized likelihood, and it became apparent through the work in Papers I and II that the prior (or penalty) on the non-stationary component in the model has a strong impact on the estimated non-stationarity structure. This is not completely unexpected since the model and the types of non-stationarity it allows is important for how the model will fill in the unobserved covariance structure. It is natural to construct a prior that shrinks the model towards stationarity, but even for stationary models there is no consensus on which prior should be used for the parameters. Progress has been made with reference priors for range (Berger et al., 2001; Paulo, 2005; Oliveira, 2007; Kazianka and Pilz, 2012; Kazianka, 2013), but these priors cannot be used as default priors in software because they are restricted to Gaussian observation noise and because they are limited to a low number of observations since they are computationally heavy. This lead us to the conclusion that we cannot develop a prior for non-stationary GRFs without first developing a prior for stationary GRFs.

Software like `INLA` (Rue et al., 2009), `spBayes` (Finley et al., 2007), `geoRGLM` (Ribeiro Jr et al., 2003) and `spTimer` (Bakar and Sahu, 2013) tend to use either uniform priors on a bounded interval, log-uniform priors on a bounded interval, inverse gamma distributions or log-Gaussian distributions on range, but these are *ad-hoc* choices without any theoretical justification. This means that within the common Bayesian hierarchical modelling framework there were no practically useful, theoretically justified priors for the spatial component in common use. We attempt to rectify this in Paper III where we develop a prior based on the Penalised Complexity framework (Simpson et al., 2014). The main result is shortly included in Simpson et al. (2014) while Paper III provides the details. The main contribution of the paper is to provide a practically useful, theoretically justified default joint prior for the range and the marginal variance for a Matérn GRF, but we also discuss an extension to non-stationary models through a formulation of the non-stationarity in the SPDE model through a change of the geometry of the base space.

The next step based on the work in Paper III is to extend the prior to non-stationarity in a less *ad-hoc* way by taking advantage of the PC prior framework. The first problem of interest is covariate-based models such as Ingebrigtsen et al. (2014a,b), but there is also interest in good priors for the models in Papers I and II. There are several ideas that have been postponed for later during the work on the thesis: the methods in Papers I and II are described for a regular grid, but could be extended to triangulations, and the method could be extended to a non-separable non-stationary spatio-temporal model through a time-derivative in the SPDE. These are all potential future directions of the work in Papers I and II.

1.6 Papers not in the thesis

There are two papers that were worked on during the Ph.D. that are not included in the thesis.

- Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2013). Non-stationary Spatial Modelling with Applications to Spatial Prediction of Precipitation. *Preprint.* arXiv:1306.0408.
- Simpson, D., Martins, T., Riebler, A., Fuglstad, G.-A., Rue, H., and Sørbye, S. (2014). Penalising model component complexity: A principled, practical approach to constructing priors. *Submitted.* arXiv:1403.4630

The first paper has been superseded by Paper II that partly includes the results of the first paper and Paper III provides the details and expands on the results for stationary GRFs found in the second paper.

References

Anderes, E. B. and Stein, M. L. (2008). Estimating deformations of isotropic Gaussian random fields on the plane. *The Annals of Statistics*, pages 719–741.

Bakar, K. S. and Sahu, S. K. (2013). spTimer: Spatio-temporal Bayesian modelling using R.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848.

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.

Bolin, D. and Lindgren, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *The Annals of Applied Statistics*, 5(1):523–550.

Bornn, L., Shaddick, G., and Zidek, J. V. (2012). Modeling nonstationary processes through dimension expansion. *Journal of the American Statistical Association*, 107(497):281–289.

Bowman, A. W. and Crujeiras, R. M. (2013). Inference for variograms. *Computational Statistics & Data Analysis*, 66:19–31.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226.

Damian, D., Sampson, P. D., and Guttorp, P. (2001). Bayesian estimation of semi-parametric non-stationary spatial covariance structures. *Environmetrics*, 12(2):161–178.

Damian, D., Sampson, P. D., and Guttorp, P. (2003). Variance modeling for nonstationary spatial processes with temporal replications. *Journal of Geophysical Research: Atmospheres*, 108(D24).

Finley, A. O., Banerjee, S., and Carlin, B. P. (2007). spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19(4):1.

Fuentes, M. (2005). A formal test for nonstationarity of spatial stochastic processes. *Journal of Multivariate Analysis*, 96(1):30–54.

Fuentes, M. (2007). Approximate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association*, 102(477):pp. 321–331.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.

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

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics*, 5:173–190. 10.1023/A:1009666805688.

Higdon, D., Swall, J., and Kern, J. (1999). Non-stationary spatial modeling. *Bayesian statistics*, 6(1):761–768.

Hu, X., Lindgren, F., Simpson, D., and Rue, H. (2013a). Multivariate Gaussian random fields with oscillating covariance functions using systems of stochastic partial differential equations. *arXiv preprint arXiv:1307.1384*.

Hu, X., Simpson, D., Lindgren, F., and Rue, H. (2013b). Multivariate Gaussian random fields using systems of stochastic partial differential equations. *arXiv preprint arXiv:1307.1379*.

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

Ingebrigtsen, R., Lindgren, F., Steinsland, I., and Martino, S. (2014b). Estimation of a non-stationary model for annual precipitation in southern norway using replicates of the spatial field. *arXiv preprint arXiv:1412.2798*.

Jun, M. and Genton, M. (2012). A test for stationarity of spatio-temporal random fields on planar and spherical domains. *Statistica Sincia*, 22:1737–1764.

Katzfuss, M. (2013). Bayesian nonstationary spatial modeling for very large datasets. *Environmetrics*, 24(3):189–200.

Kazianka, H. (2013). Objective Bayesian analysis of geometrically anisotropic spatial data. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(4):514–537.

Kazianka, H. and Pilz, J. (2012). Objective Bayesian analysis of spatial data with uncertain nugget and range parameters. *Canadian Journal of Statistics*, 40(2):304–327.

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

Matérn, B. et al. (1960). Spatial variation. stochastic models and their application to some problems in forest surveys and other sampling investigations. *Meddelanden fran statens Skogsforskningsinstitut*, 49(5).

Neto, J. H. V., Schmidt, A. M., and Guttorp, P. (2014). Accounting for spatially varying directional effects in spatial covariance structures. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(1):103–122.

Oliveira, V. d. (2007). Objective Bayesian analysis of spatial data with measurement error. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 35(2):pp. 283–301.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506.

Paulo, R. (2005). Default priors for Gaussian processes. *The Annals of Statistics*, 33(2):556–582.

Ribeiro Jr, P. J., Christensen, O. F., and Diggle, P. J. (2003). geoR and geoRglm: software for model-based geostatistics. In *Proceedings of DSC*, page 2.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

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

Schmidt, A. M., Guttorp, P., and O'Hagan, A. (2011). Considering covariates in the covariance structure of spatial processes. *Environmetrics*, 22(4):487–500.

Schmidt, A. M. and O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):743–758.

Simpson, D., Lindgren, F., and Rue, H. (2012). In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics*, 23(1):65–74.

Simpson, D. P., Martins, T. G., Riebler, A., Rue, H., Fuglstad, G.-A., and Sørbye, S. H. (2014). Penalising model component complexity: A principled, practical approach to constructing priors. *arXiv preprint arXiv:1403.4630*.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Stein, M. L., Chi, Z., and Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):275–296.

Sun, Y., Li, B., and Genton, M. G. (2012). Geostatistics for large datasets. In Porcu, E., Montero, J., Schlather, M., Bickel, P., Diggle, P., Fienberg, S., Krickeberg, K., Olkin, I., Wermuth, N., and Zeger, S., editors, *Advances and Challenges in Space-time Modelling of Natural Events*, volume 207 of *Lecture Notes in Statistics*, pages 55–77. Springer Berlin Heidelberg. 10.1007/978-3-642-17086-7_3.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41(3/4):pp. 434–449.

Whittle, P. (1963). Stochastic-processes in several dimensions. *Bulletin of the International Statistical Institute*, 40(2):974–994.

Paper I

Exploring a New Class of Non-stationary Spatial Gaussian Random Fields with Varying Local Anisotropy

Fuglstad, G.-A., Lindgren, F., Simpson, D., and Rue, H.

Statistica Sinica 25(1):115–133, 2015.

Exploring a New Class of Non-stationary Spatial Gaussian Random Fields with Varying Local Anisotropy

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

[1]Department of Mathematical Sciences, NTNU, Norway
[2]Department of Mathematical Sciences, University of Bath, United Kingdom

Abstract

Gaussian random fields (GRFs) play an important part in spatial modelling, but can be computationally infeasible for general covariance structures. An efficient approach is to specify GRFs via stochastic partial differential equations (SPDEs) and derive Gaussian Markov random field (GMRF) approximations of the solutions. We consider the construction of a class of non-stationary GRFs with varying local anisotropy, where the local anisotropy is introduced by allowing the coefficients in the SPDE to vary with position. This is done by using a form of diffusion equation driven by Gaussian white noise with a spatially varying diffusion matrix. This allows for the introduction of parameters that control the GRF by parametrizing the diffusion matrix. These parameters and the GRF may be considered to be part of a hierarchical model and the parameters estimated in a Bayesian framework. The results show that the use of an SPDE with non-constant coefficients is a promising way of creating non-stationary spatial GMRFs that allow for physical interpretability of the parameters, although there are several remaining challenges that would need to be solved before these models can be put to general practical use.

Keywords: Non-stationary, Spatial, Gaussian random fields, Gaussian Markov random fields, Anisotropy, Bayesian

1 Introduction

Many spatial models for continuously indexed phenomena, such as temperature, precipitation and air pollution, are based on Gaussian random fields (GRFs). This is mainly due to the fact that their theoretical properties are well understood and that their distributions can be fully described by mean and covariance functions. In principle, it is enough to specify the mean at each location and the covariance between any two locations. However, specifying covariance functions is hard and specifying covariance functions that can be controlled by parameters in useful ways is even harder. This is the reason why the covariance function usually is selected from a class of known covariance functions such as the exponential covariance function, the Gaussian covariance function, or the Matérn covariance function.

But even when the covariance function is selected from one of these classes, the feasible problem sizes are severely limited by a cubic increase in computation time as a function of the number of observations and a quadratic increase in computation time as a function of the number of prediction locations. This computational challenge is usually tackled either by reducing the dimensionality of the problem (Cressie and Johannesson (2008); Banerjee, Gelfand, Finley, and Sang (2008)), by introducing sparsity in the precision matrix (Rue and Held (2005)) or the covariance matrix (Furrer, Genton, and Nychka (2006)), or by using an approximate likelihood (Stein, Chi, and Welty (2004); Fuentes (2007)). Sun, Li, and Genton (2012) offers comparisons of the advantages and challenges associated with the usual approaches to large spatial datasets.

We explore a new class of non-stationary GRFs that provide both an easy way to specify the parameters and allow for fast computations. The main computational tool used is Gaussian Markov random fields (GMRFs) (Rue and Held (2005)) with a spatial Markovian structure where each position is conditionally dependent only on positions close to itself. The strong connection between the Markovian structure and the precision matrix results in sparse precision matrices that can be exploited in computations. The main problem associated with such an approach is that GMRFs must be constructed through conditional distributions, which presents a challenge as it is generally not easy to determine whether a set

of conditional distributions gives a valid joint distribution. Additionally, the conditional distributions have to be controlled by useful parameters in such a way that not only is the joint distribution valid, but also that the effect of the parameters is understood. Lastly, it is desirable that the GMRF is a consistent approximation of a GRF in the sense that when the distances between the positions decrease, the GMRF "approaches" a continuous GRF. These issues are even more challenging for non-stationary GMRFs. It is is extremely hard to specify the non-stationarity directly through conditional distributions.

There is no generally accepted way to handle non-stationary GRFs, but many approaches have been suggested. There is a large literature on methods based on the deformation method of Sampson and Guttorp (1992), where a stationary process is made non-stationary by deforming the space on which it is defined. Several Bayesian extensions of the method have been proposed (Damian, Sampson, and Guttorp (2001, 2003); Schmidt and O'Hagan (2003); Schmidt, Guttorp, and O'Hagan (2011)), but all these methods require replicated realizations which might not be available. There has been some development towards an approach for a single realization, but with a "densely" observed realization (Anderes and Stein (2008)). Other approaches use kernels which are convolved with Gaussian white noise (Higdon (1998); Paciorek and Schervish (2006)), weighted sums of stationary processes (Fuentes (2001)) and expansions into a basis such as a wavelet basis (Nychka, Wikle, and Royle (2002)). Conceptually simpler methods have been made with "stationary windows" (Haas (1990b,a)) and with piecewise stationary Gaussian processes (Kim, Mallick, and Holmes (2005)). There has also been some progress with methods based on the spectrum of the processes (Fuentes (2001, 2002a,b)). Recently, a new type of method based on a connection between stochastic partial differential equations (SPDEs) and some classes of GRFs was proposed by Lindgren, Rue, and Lindström (2011). They use an SPDE to model the GRF and construct a GMRF approximation to the GRF for computations. An application of a non-stationary model of this type to ozone data can be found in Bolin and Lindgren (2011) and an application to precipitation data can be found in Ingebrigtsen, Lindgren, and Steinsland (2013).

This paper extends on the work of Lindgren, Rue, and Lindström (2011) and explores the possibility of constructing a non-stationary GRF

by varying the local anisotropy. The interest lies both in considering the different types of structures that can be achieved, and how to parametrize the GRF and estimate the parameters in a Bayesian setting. The construction of the GRF is based on an SPDE that describes the GRF as the result of a linear filter applied to Gaussian white noise. Basically, the SPDE expresses how the smoothing of the Gaussian white noise varies at different locations. This construction bears some resemblance to the deformation method of Sampson and Guttorp (1992) in the sense that parts of the spatial variation of the linear filter can be understood as a local deformation of the space, only with an associated spatially varying variance for the Gaussian white noise. The main idea for computations is that since this filter works locally, it implies a Markovian structure on the GRF. This Markovian structure can be transferred to a GMRF which approximates the GRF, and in turn fast computations can be done with sparse matrices.

This paper presents a new type of model and the main goal is to explore what can be achieved in terms of models and inference with the model. Section 2 contains the motivation and introduction to the class of non-stationary GRFs that is studied in the other sections. The form of the SPDE that generates the class is given and it is related to more standard constructions of GMRFs. In Section 3 illustrative examples are given on both stationary and non-stationary constructions. This includes some discussion on how to control the non-stationarity of the GRF. Then in Section 4 we discuss parameter estimation for these types of models. The paper ends with discussion of extensions in Section 5, and general discussion and concluding remarks in Section 6.

2 New class of non-stationary GRFs

A GMRF $\boldsymbol{u}$ is usually parametrized through a mean $\boldsymbol{\mu}$ and a precision matrix $\mathbf{Q}$ such that $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1})$. The main advantage of this formulation is that the Markovian structure is represented in the non-zero structure of the precision matrix $\mathbf{Q}$ (Rue and Held (2005)). Off-diagonal entries are non-zero if and only if the corresponding elements of $\boldsymbol{u}$ are conditionally independent. This can be seen from the conditional properties

of a GMRF,

$$\mathrm{E}(u_i|\boldsymbol{u}_{-i}) = \mu_i - \frac{1}{Q_{i,i}} \sum_{j \neq i} Q_{i,j}(u_j - \mu_j), \qquad \mathrm{Var}(u_i|\boldsymbol{u}_{-i}) = \frac{1}{Q_{i,i}},$$

where $\boldsymbol{u}_{-i}$ denotes the vector $\boldsymbol{u}$ with element i deleted. For a spatial GMRF the non-zeros of $\mathbf{Q}$ can correspond to grid-cells that are close to each other in a grid, neighbouring regions in a Besag model, and so on. However, even when this non-zero structure is determined it is not clear what values should be given to the non-zero elements of the precision matrix. This is the framework of the conditionally auto-regressive (CAR) models, whose conception predates the advances in modern computational statistics (Whittle (1954); Besag (1974)). In the multivariate Gaussian case it is clear that the requirement for a valid joint distribution is that $\mathbf{Q}$ be positive definite, not an easy condition to check.

Specification of a GMRF through its conditional properties is usually done in a somewhat ad-hoc manner. For regular grids, a process such as random walk can be constructed and the only major issue is to get the conditional variance correct as a function of step-length. For irregular grids the situation is not as clear because the conditional means and variances must depend on the varying step-lengths. In Lindgren and Rue (2008) it is demonstrated that some such constructions for second-order random walk can lead to inconsistencies as new grid points are added, and they offer a surprisingly simple construction for second-order random walk based on the SPDE

$$-\frac{\partial^2}{\partial x^2} u(x) = \sigma \mathcal{W}(x),$$

where $\sigma > 0$ and $\mathcal{W}$ is standard Gaussian white noise. If the precision matrix is chosen according to their scheme one does not have to worry about scaling as the grid is refined, as it automatically approaches the continuous second-order random walk. There is an automatic procedure to select the form of the conditional means and variances.

A one-dimensional second-order random walk is a relatively simple example of a process with the same behaviour everywhere. To approximate a two-dimensional, non-stationary GRF, a scheme would require (possibly) different anisotropy and correct conditional variance at each location. To select the precision matrix in this situation poses a large problem and

there is abundant use of simple models such as a spatial moving average

$$\mathrm{E}(u_{i,j}|\boldsymbol{u}_{-\{(i,j)\}}) = \frac{1}{4}(u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1})$$

with a constant conditional variance $1/\alpha$. There are ad-hoc ways to extend such a scheme to a situation with varying step-lengths in each direction, but little theory for more irregular choices of locations.

We start with the close connection between SPDEs and some classes of GRFs as presented in Lindgren, Rue, and Lindström (2011) that is not plagued by such issues. From Whittle (1954), it is known that the SPDE

$$(\kappa^2 - \Delta)u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \qquad \boldsymbol{s} \in \mathbb{R}^2, \tag{2.1}$$

where $\kappa^2 > 0$ and $\Delta = \frac{\partial^2}{\partial s_1^2} + \frac{\partial^2}{\partial s_2^2}$ is the Laplacian, gives rise to a GRF u with the Matérn covariance function

$$r(\boldsymbol{s}) = \frac{1}{4\pi\kappa^2}(\kappa||\boldsymbol{s}||)K_1(\kappa||\boldsymbol{s}||),$$

where K_1 is the modified Bessel function of the second kind of order 1. Equation (2.1) can be extended to fractional operator orders in order to obtain other smoothness parameters in the Matérn covariance function. However, for practical applications, the true smoothness of the field is very hard to estimate from data, in particular when the model is used in combination with an observation noise model. Restricting the development to smoothness 1 in the Matérn family is therefore unlikely to be a major practical serious limitation. For practical computations the model is discretized using methods similar to those in Lindgren, Rue, and Lindström (2011), which does permit other operator orders. Integer orders are easiest but, for stationary models, fractional orders are also achievable (Lindgren, Rue, and Lindström (2011, Authors' discussion response)). For non-stationary models, techniques similar to Bolin (2013, Section 4.2) are possible. This means that even though we here restrict the model development to the special case in (2.1), other smoothnesses, e.g. exponential covariances, are reachable by combining the different approximation techniques.

The intriguing part, that Lindgren, Rue, and Lindström (2011) expanded upon in (2.1), is that $(\kappa^2 - \Delta)$ can be interpreted as a linear

filter acting locally. This means that if the continuously indexed process u were instead represented by a GMRF $\boldsymbol{u}$ on a grid or a triangulation, with appropriate boundary conditions, one could replace this operator with a matrix, say $\mathbf{B}(\kappa^2)$, only involving neighbours of each location such that (2.1) becomes approximately

$$\mathbf{B}(\kappa^2)\boldsymbol{u} \sim \mathcal{N}(0, \mathbf{I}). \tag{2.2}$$

The matrix $\mathbf{B}(\kappa^2)$ depends on the chosen grid but, after the relationship is derived, the calculation of $\mathbf{B}(\kappa^2)$ is straightforward for any κ^2. Since $\mathbf{B}(\kappa^2)$ is sparse, the resulting precision matrix $\mathbf{Q}(\kappa^2) = \mathbf{B}(\kappa^2)^{\mathrm{T}}\mathbf{B}(\kappa^2)$ for $\boldsymbol{u}$ is also sparse. This means that by correctly discretizing the operator (or linear filter), it is possible to devise a GMRF with approximately the same distribution as the continuously indexed GRF. And because it comes from a continuous equation one does not have to worry about changing behaviour as the grid is refined.

The class of models that are studied in this paper is the one that can be constructed from (2.1), but with anisotropy added to the Δ operator. A function $\mathbf{H}$, that gives 2×2 symmetric positive definite matrices at each position, is introduced and the operator is changed to

$$\begin{aligned}\nabla \cdot \mathbf{H}(\boldsymbol{s})\nabla = &\frac{\partial}{\partial s_1}\left(h_{11}(\boldsymbol{s})\frac{\partial}{\partial s_1}\right) + \frac{\partial}{\partial s_1}\left(h_{12}(\boldsymbol{s})\frac{\partial}{\partial s_2}\right) \\ &+ \frac{\partial}{\partial s_2}\left(h_{21}(\boldsymbol{s})\frac{\partial}{\partial s_1}\right) + \frac{\partial}{\partial s_2}\left(h_{22}(\boldsymbol{s})\frac{\partial}{\partial s_2}\right).\end{aligned}$$

This induces different strengths of local dependence in different directions, which results in a range that varies with direction at all locations. Further, it is necessary for the discretization procedure to restrict the SPDE to a bounded domain. The chosen SPDE is

$$(\kappa^2 - \nabla \cdot \mathbf{H}(\boldsymbol{s})\nabla)u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \qquad \boldsymbol{s} \in \mathcal{D} = [A_1, B_1] \times [A_2, B_2] \subset \mathbb{R}^2, \tag{2.3}$$

where the rectangular domain makes it possible to use periodic boundary conditions. Neither the rectangular shape of the domain nor the periodic boundary conditions are essential restrictions for the model, but are merely the practical restrictions we choose to work with, in order to focus on the non-stationarity itself.

When using periodic boundary conditions when approximating the likelihood of a stationary process on an unbounded domain, the parameter estimates are biased, e.g., when using the Whittle likelihood in the two-dimensional case (Dahlhaus and Künsch (1987)). However, as Lindgren, Rue, and Lindström (2011, Appendix A.4) notes for the case with Neumann boundary conditions, normal derivatives set to zero, the effect of the boundary conditions is limited to a region in the vicinity of the boundary. At a distance greater than twice the correlation range away from the boundary the bounded domain model is nearly indistinguishable from the model on an unbounded domain. Therefore, the bias due to boundary effects can be eliminated by embedding the domain of interest into a larger region, in effect moving the boundary away from where it would influence the likelihood function. For non-stationary models, defining appropriate boundary conditions becomes part of the practical model formulation itself. For simplicity we ignore this issue here, leaving boundary specification for future development, but provide some additional practical comments in Section 5.

Both for interpretation and the practical use of (2.3) it is useful to decompose $\mathbf{H}$ into scalar functions. The anisotropy due to $\mathbf{H}$ is decomposed as $\mathbf{H}(\boldsymbol{s}) = \gamma\mathbf{I}_2 + \boldsymbol{v}(\boldsymbol{s})\boldsymbol{v}(\boldsymbol{s})^{\mathrm{T}}$, where γ specifies the isotropic, baseline effect, and the vector field $\boldsymbol{v}(\boldsymbol{s}) = [v_x(\boldsymbol{s}), v_y(\boldsymbol{s})]^{\mathrm{T}}$ specifies the direction and magnitude of the local, extra anisotropic effect at each location. In this way, one can, loosely speaking, think of different Matérn-like fields locally each with its own anisotropy that are combined into a full process. An example of an extreme case of a process with a strong local anisotropic effect is shown in Example 3.2. The example shows that there is a close connection between the vector field and the resulting covariance structure of the GRF.

The main computational challenge is to determine the appropriate discretization of the SPDE in (2.3), that is how to derive a matrix $\mathbf{B}$ such as in (2.2). The idea is to look to the field of numerics for discretization methods for differential equations, then to combine these with properties of Gaussian white noise. We use that for a Lebesgue measurable subset A of $\mathbb{R}^n$, for some $n > 0$,

$$\int_A \mathcal{W}(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s} \sim \mathcal{N}(0, |A|),$$

where $|A|$ is the Lebesgue measure of A, and for two disjoint Lebesgue measurable subsets A and B of $\mathbb{R}^n$ the integral over A and the integral over B are independent (Adler and Taylor (2007, pp. 24–25)). A matrix equation such as (2.2) was derived for the SPDE in (2.3) with a finite volume method. The derivations are quite involved and technical and are given in a supplementary document. However, when the form of the discretized SPDE has been derived as an expression of the coefficients in the SPDE and the grid, the conversion from SPDE to GMRF is automatic for any choice of coefficients and rectangular domain.

3 Examples of models

The simplest case of (2.3) is with constant coefficients. In this case one has an isotropic model (up to boundary effects) if $\mathbf{H}$ is a constant times the identity matrix or a stationary anisotropic model (up to boundary effects) if this is not the case. In both cases it is possible to calculate an exact expression for the covariance function and the marginal variance for the corresponding SPDE solved over $\mathbb{R}^2$.

For this purpose write

$$\mathbf{H} = \begin{bmatrix} H_1 & H_2 \\ H_2 & H_3 \end{bmatrix},$$

where H_1, H_2, and H_3 are constants. This gives the SPDE

$$\left[\kappa^2 - H_1\frac{\partial^2}{\partial x^2} - 2H_2\frac{\partial^2}{\partial x \partial y} - H_3\frac{\partial^2}{\partial y^2}\right] u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \qquad \boldsymbol{s} \in \mathbb{R}^2.$$

If λ_1 and λ_2 are the eigenvalues of $\mathbf{H}$, then the solution of the SPDE is actually only a rotated version of the solutions of

$$\left[\kappa^2 - \lambda_1\frac{\partial^2}{\partial \tilde{x}^2} - \lambda_2\frac{\partial^2}{\partial \tilde{y}^2}\right] u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \qquad \boldsymbol{s} \in \mathbb{R}^2. \tag{3.1}$$

Here the new x-axis is parallel to the eigenvector of $\mathbf{H}$ corresponding to λ_1 in the old coordinate system and the new y-axis is parallel to the eigenvector of $\mathbf{H}$ corresponding to λ_2.

The marginal variance of u is

$$\sigma_m^2 = \frac{1}{4\pi\kappa^2\sqrt{\det(\mathbf{H})}} = \frac{1}{4\pi\kappa^2\sqrt{\lambda_1\lambda_2}}.$$

A proof is given in the supplementary material. One can think of the eigenvectors of $\mathbf{H}$ as the two principal directions and λ_1 and λ_2 as a measure of the "strength" of the diffusion in these principal directions. Additionally, if $\lambda_1 = \lambda_2$, which is equivalent to $\mathbf{H}$ being equal to a constant times the identity matrix, the SPDE is rotation and translation invariant and the solution is isotropic. If $\lambda_1 \neq \lambda_2$, the SPDE is still translation invariant, but not rotation invariant, and the solutions are stationary, but not isotropic.

In our case the domain is not $\mathbb{R}^2$, but $[0, A] \times [0, B]$ with periodic boundary conditions. This means that a boundary effect is introduced and the above results are only approximately true.

3.1 Stationary models

For a constant $\mathbf{H}$ the SPDE in (2.3) becomes

$$[\kappa^2 - \nabla \cdot \mathbf{H}\nabla]u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \qquad \boldsymbol{s} \in [0, A] \times [0, B].$$

This SPDE can be rewritten as

$$[1 - \nabla \cdot \hat{\mathbf{H}}\nabla]u(\boldsymbol{s}) = \sigma\mathcal{W}(\boldsymbol{s}), \qquad \boldsymbol{s} \in [0, A] \times [0, B], \tag{3.2}$$

where $\hat{\mathbf{H}} = \mathbf{H}/\kappa^2$ and $\sigma = 1/\kappa^2$. From this form it is clear that σ is only a scale parameter and that it is enough to solve for $\sigma = 1$ and then multiply the solution with the desired value of σ. Therefore, it is the effect of $\hat{\mathbf{H}}$ that is most interesting to study.

It is useful to parametrize $\hat{\mathbf{H}}$ as $\hat{\mathbf{H}} = \gamma\mathbf{I}_2 + \beta\boldsymbol{v}(\theta)\boldsymbol{v}(\theta)^{\mathrm{T}}$, where $\boldsymbol{v}(\theta) = [\cos(\theta), \sin(\theta)]^{\mathrm{T}}$, $\gamma > 0$, and $\beta > 0$. In this parametrization one can think of γ as the coefficient of the second order derivative in the direction orthogonal to $\boldsymbol{v}(\theta)$ and $\gamma+\beta$ as the coefficient of the second order derivative in the direction $\boldsymbol{v}(\theta)$. Ignoring boundary effects, γ and $\gamma + \beta$ are the coefficients of the second order derivatives in (3.1) and θ is how much the coordinate system has been rotated in the positive direction.

Example 3.1 (Stationary GMRF)**.** Here we consider the effects of using a constant $\hat{\mathbf{H}}$. Use the SPDE in (3.2) with domain $[0, 20] \times [0, 20]$ and periodic boundary conditions, and discretize with a regular 200×200 grid. Two different values of $\hat{\mathbf{H}}$ are used, an isotropic case with $\hat{\mathbf{H}} = \mathbf{I}_2$ and an anisotropic case with $\gamma = 1$, $\beta = 8$, and $\theta = \pi/4$. The anisotropic case corresponds to a coefficient 9 in the x-direction and a coefficient 1 in the y-direction, and then a rotation of $\pi/4$ in the positive direction. The isotropic GMRF has marginal variances 0.0802 and the anisotropic GMRF has marginal variances 0.0263.

Figure 1 shows one realization for each of the cases. Comparing Figure 1(a) and Figure 1(b) it seems that the direction with the higher coefficient for the second-order derivative has longer range and more regular behaviour. Compared to the corresponding partial differential equation (PDE) without the white noise, this is what one would expect since large values of the coefficient penalize large values of the second order derivatives.

One expects that the correlation range increases when the coefficient is increased. This is in fact what happens. Figure 2 shows the correlation of the variable at $(9.95, 9.95)$ with every other point in the grid for the isotropic and the anisotropic case. This is sufficient to describe all the correlations since the solutions are stationary. One can see that the iso-correlation curves are close to ellipses with semi-axes along $\boldsymbol{v}(\theta)$ and the direction orthogonal to $\boldsymbol{v}(\theta)$, and that the correlation decreases most slowly and most quickly in the directions used to specify $\hat{\mathbf{H}}$, with slowest decrease along $\boldsymbol{v}(\theta)$. It is interesting that the isotropic and the non-isotropic cases have approximately the same length for the minor semi-axis of the iso-correlation curves, and that the major semi-axis is longer for the anisotropic case, lengths being connected with $\sqrt{\gamma}$ and $\sqrt{\gamma + \beta}$.

The use of 3 parameters thus allows for the creation of GMRFs that are more regular in one direction than the other. One can use the parameters γ, β, and θ to control the form of the correlation function, and σ to get the desired marginal variance.

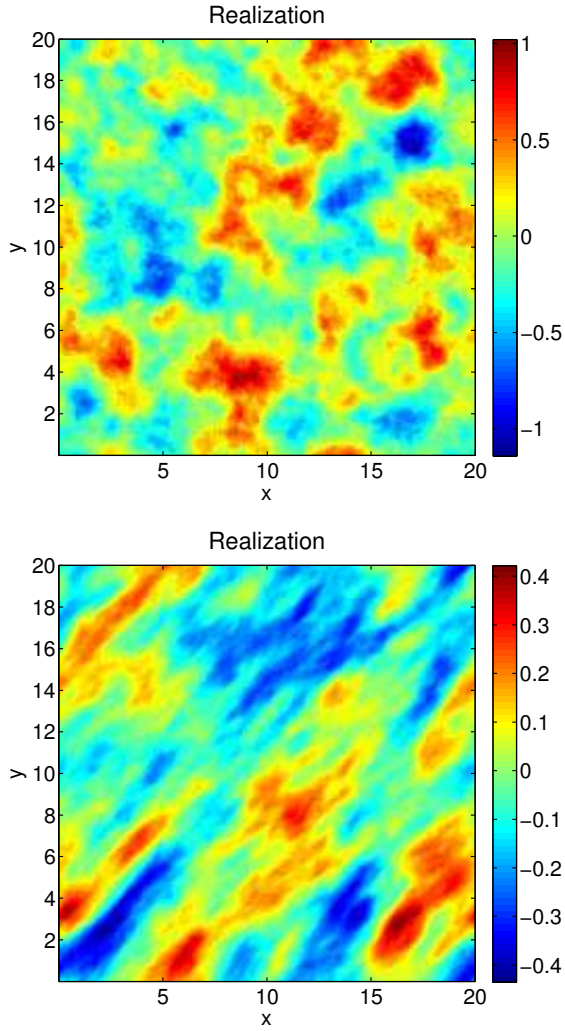


Figure 1: (a) Realization from the SPDE in Example 3.1 on $[0, 20]^2$ with a 200×200 grid and periodic boundary conditions with $\gamma = 1$, $\beta = 0$, and $\theta = 0$. (b) Realization from the SPDE in Example 3.1 on $[0, 20]^2$ with a 200×200 grid and periodic boundary conditions with $\gamma = 1$, $\beta = 8$, and $\theta = \pi/4$.

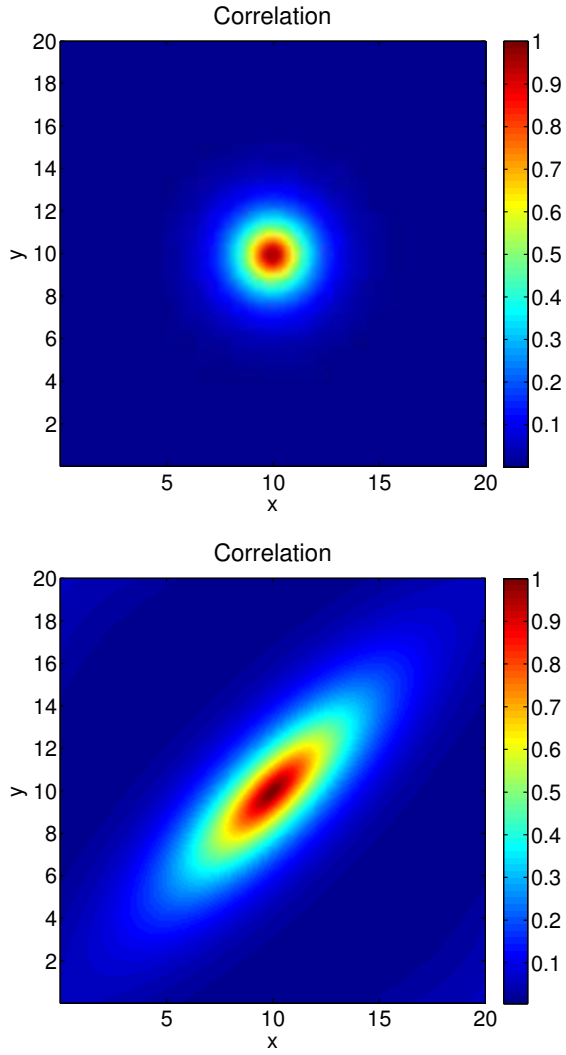


Figure 2: (a) Correlation of the centre with all other points for the solution of the SPDE in Example 3.1 on $[0, 20]^2$ with a 200×200 grid and periodic boundary conditions with $\gamma = 1$, $\beta = 0$, and $\theta = 0$. (b) Correlation of the centre with all other points for the SPDE in Example 3.1 on $[0, 20]^2$ with a 200×200 grid and periodic boundary conditions with $\gamma = 1$, $\beta = 8$, $\theta = \pi/4$.

3.2 Non-stationary models

To make the solution of the SPDE in (2.3) non-stationary, either κ^2 or $\mathbf{H}$ has to be a non-constant function. One way to achieve non-stationarity is by choosing $\mathbf{H}(\boldsymbol{s}) = \gamma\mathbf{I}_2 + \beta\boldsymbol{v}(\boldsymbol{s})\boldsymbol{v}(\boldsymbol{s})^{\mathrm{T}}$, where $\boldsymbol{v}$ is a non-constant vector field on $[0, A]\times[0, B]$ satisfying the periodic boundary conditions and $\gamma > 0$ and $\beta > 0$ are constants.

Example 3.2 (Non-stationary GMRF)**.** Use the domain $[0, 20]^2$ with a 200×200 grid and periodic boundary conditions for the SPDE in (2.3). Let κ^2 be 1 and let $\mathbf{H}$ be given as $\mathbf{H}(\boldsymbol{s}) = \gamma\mathbf{I}_2 + \beta\boldsymbol{v}(\boldsymbol{s})\boldsymbol{v}(\boldsymbol{s})^{\mathrm{T}}$, where $\boldsymbol{v}$ is a 2-dimensional vector field on $[0, 20]^2$ which satisfies the periodic boundary conditions and $\gamma > 0$ and $\beta > 0$ are constants.

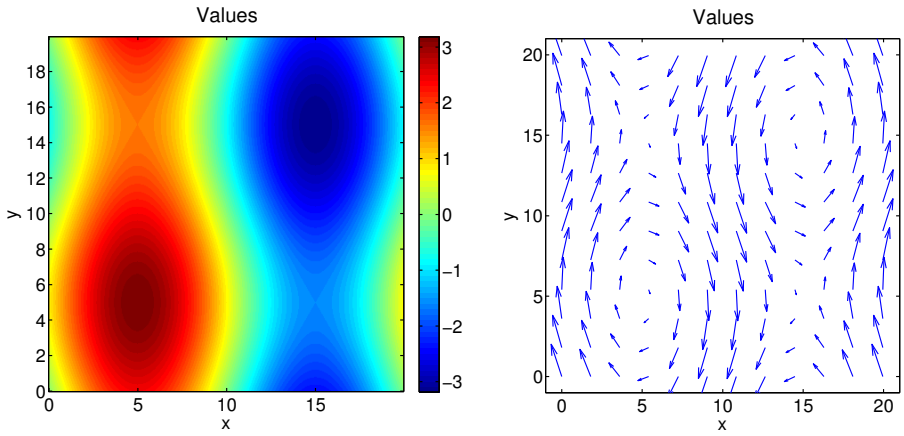
To create an interesting vector field, start with $f : [0, 20]^2 \to \mathbb{R}$ defined by

$$f(x, y) = \left(\frac{10}{\pi}\right)\left(\frac{3}{4}\sin(2\pi x/20) + \frac{1}{4}\sin(2\pi y/20)\right).$$

Then calculate the gradient ∇f and let $\boldsymbol{v} : [0, 20]^2 \to \mathbb{R}^2$ be the gradient rotated 90° counter-clockwise at each point. Figure 3(a) shows the values of the function f and Figure 3(b) shows the resulting vector field $\boldsymbol{v}$. The vector field is calculated on a 400×400 regular grid, because the values between neighbouring cells in the discretization are needed.

Figure 4(a) shows one realization from the resulting GMRF with $\gamma = 0.1$ and $\beta = 25$. A much higher value for β than γ is chosen to illustrate the connection between the vector field and the resulting covariance structure. From the realization it is clear that there is stronger dependence along the directions of the vector field shown in Figure 3(b) at each point than in the other directions. In addition, from Figure 4(b) it seems that positions with large values for the norm of the vector field have smaller marginal variance than positions with small values, and vice versa. This feature introduces an undesired connection between anisotropy and marginal variances. It is possible to reduce this interaction by reformulating the controlling SPDE, as discussed briefly in Section 5.

From Figure 5 one can see that the correlations depend on the direction and norm of the vector field, and that there is clearly non-stationarity. Figure 5(a) and Figure 5(c) show that the correlations with the positions $(4.95, 1.95)$ $(4.95, 7.95)$ tend to follow the vector field around the point



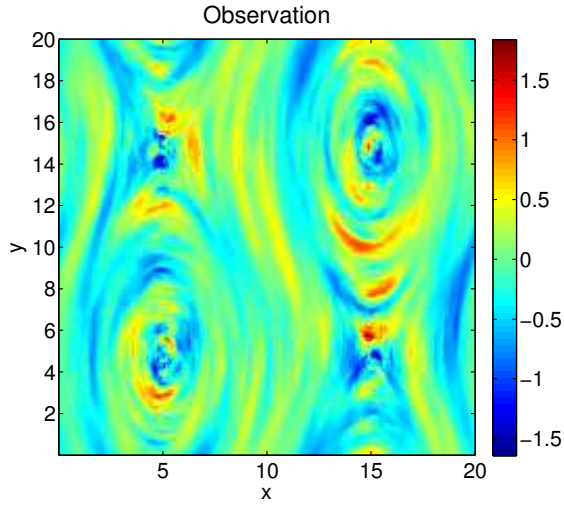
(a) The function used to create the vector field.

(b) The resulting vector field.

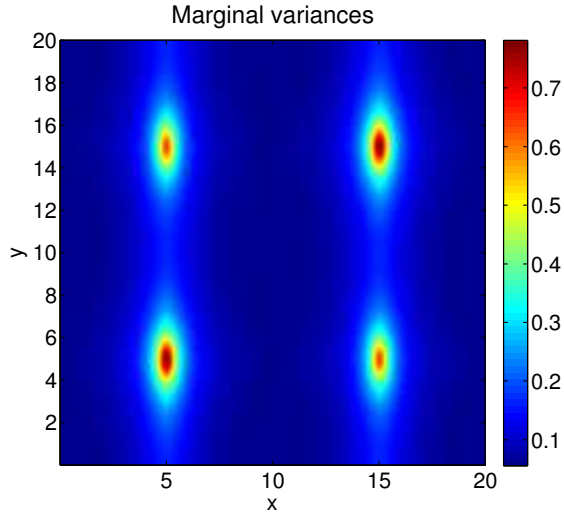
Figure 3: The gradient of the function illustrated in (a) is calculated and rotated 90° counter-clockwise at each point to give the vector field illustrated in (b).

$(5, 5)$, whereas Figure 5(b) and Figure 5(d) show that the correlations with the positions $(14.95, 1.95)$ and $(14.95, 7.95)$ tend to follow the vector field away from the point $(15, 5)$. Figure 5(e) shows that the correlations with position $(4.95, 4.95)$ and every other point are not isotropic, but concentrated close to the point itself, and Figure 5(f) shows that the correlations with position $(14.95, 4.95)$ have high correlation along four directions that extend out from the point.

We see that allowing $\mathbf{H}$ to be non-constant can vary the dependence structure in more interesting ways than in stationary anisotropic fields. Using a vector field to control how $\mathbf{H}$ varies means that the resulting correlation structure can be partially visualized from the vector field. When $\gamma > 0$ this construction guarantees that $\mathbf{H}$ is everywhere positive definite.



(a) One realization.



(b) Marginal variances.

Figure 4: One observation and the marginal variances of the solution of the SPDE in Equation (2.3) on a 200×200 regular grid of $[0, 20]^2$ with periodic boundary conditions, $\kappa^2 \equiv 1$ and $\mathbf{H} = 0.1\mathbf{I}_2 + 25\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}$, where $\boldsymbol{v}$ is the vector field described in Example 3.2.

4 Inference

4.1 Posterior distribution and parametrization

For inference, we introduce parameters that control the behaviour of the coefficients in (2.3) and, in turn, the behaviour of the GMRF. This is done by expanding each of the functions in a basis and using a linear combination of the basis functions weighted by parameters. For κ^2 only one parameter, say θ_1, is needed as it is assumed constant, but for the function $\mathbf{H}$ a vector of parameters $\boldsymbol{\theta}_2$ is needed. Set $\boldsymbol{\theta} = (\theta_1, \boldsymbol{\theta}_2^{\mathrm{T}})$ and give it a prior $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$. Then for each value of $\boldsymbol{\theta}$, a discretization, described in the supplementary document, is used to construct the GMRF $\boldsymbol{u}|\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}(\boldsymbol{\theta})^{-1})$. Combine the prior of $\boldsymbol{\theta}$ with this conditional distribution to find the joint distribution of the parameters and $\boldsymbol{u}$. With a model for how an observation $\mathbf{y}$ is made from the underlying GMRF, one forms a hierarchical spatial model. The relationship between $\boldsymbol{y}$ and $\boldsymbol{u}$ is chosen to be particularly simple, namely that linear combinations of $\mathbf{u}$ are observed with Gaussian noise, $\boldsymbol{y}|\boldsymbol{u} \sim \mathcal{N}(\mathbf{A}\boldsymbol{u}, \mathbf{Q}_{\mathrm{N}}^{-1})$, where $\mathbf{Q}_{\mathrm{N}}$ is a known precision matrix.

The purpose of the hierarchical model is to do inference on $\boldsymbol{\theta}$ based on an observation of $\boldsymbol{y}$. With a Gaussian latent model it is possible to integrate out the latent field $\boldsymbol{u}$ exactly and this leads to the log-posterior

$$\begin{aligned}\log(\pi(\boldsymbol{\theta}|\boldsymbol{y})) = \\ \quad \text{Const} + \log(\pi(\boldsymbol{\theta})) + \frac{1}{2}\log(|\mathbf{Q}(\boldsymbol{\theta})|) \\ \quad - \frac{1}{2}\log(|\mathbf{Q}_{\mathrm{C}}(\boldsymbol{\theta})|) + \frac{1}{2}\boldsymbol{\mu}_C(\boldsymbol{\theta})^{\mathrm{T}}\mathbf{Q}_{\mathrm{C}}(\boldsymbol{\theta})\boldsymbol{\mu}_C(\boldsymbol{\theta}),\end{aligned} \tag{4.1}$$

where $\mathbf{Q}_{\mathrm{C}}(\boldsymbol{\theta}) = \mathbf{Q}(\boldsymbol{\theta}) + \mathbf{A}^{\mathrm{T}}\mathbf{Q}_{\mathrm{N}}\mathbf{A}$ and $\boldsymbol{\mu}_{\mathrm{C}}(\boldsymbol{\theta}) = \mathbf{Q}_{\mathrm{C}}(\boldsymbol{\theta})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{Q}_{\mathrm{N}}\boldsymbol{y}$. In (4.1) one sees that the posterior distribution of $\boldsymbol{\theta}$ contains terms that are hard to handle analytically. It is hard to say anything about both the determinants and the quadratic term as functions of $\boldsymbol{\theta}$. Therefore, the inference is done numerically. The model is of a form that could be handled by the INLA methodology (Rue, Martino, and Chopin (2009)), but when this was written, the R-INLA software[1] did not have the model implemented.

[1] `www.r-inla.org`

Instead the parameters are estimated with maximum a posteriori estimates based on the posterior density given in (4.1). Standard deviations are estimated from the square roots of the diagonal elements of the observed information matrix.

The parametrization of $\mathbf{H}$ in the previous section employs a pre-defined vector field together with a parameter β that controls the magnitude of the anisotropy due to this vector field. This is a useful representation for achieving a desired dependence structure, but in a inference setting there may not be any pre-defined vector field. Therefore, the vector field itself must be estimated. In this context the decomposition, $\mathbf{H}(\boldsymbol{s}) = \gamma \mathbf{I}_2 + \boldsymbol{v}(\boldsymbol{s})\boldsymbol{v}(\boldsymbol{s})^{\mathrm{T}}$, is more useful. For inference it is necessary to control the vector field by a finite number of parameters. The simple case of a constant matrix requires 3 parameters. Use parameters γ, v_1, and v_2 and write

$$\mathbf{H}(\boldsymbol{s}) \equiv \gamma \mathbf{I}_2 + \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \begin{bmatrix} v_1 & v_2 \end{bmatrix}.$$

If $\mathbf{H}$ is not constant, it is necessary to parametrize the vector field $\boldsymbol{v}$ in some manner. Any vector field is possible for $\boldsymbol{v}$, so a basis that can generate any vector field is desirable. The Fourier basis possesses this property, but is only one of many possible choices. Let the domain be $[0, A] \times [0, B]$ and assume that $\boldsymbol{v}$ is a differentiable, periodic vector field on the domain. Then each component of the vector field can be written as a Fourier series of the form

$$\sum_{(k,l) \in \mathbb{Z}^2} C_{k,l} \exp\left[2\pi i \left(\frac{k}{A}x + \frac{l}{B}y\right)\right],$$

where i is the imaginary unit. But since the components are real-valued, each of them can also be written as a real 2-dimensional Fourier series of the form

$$A_{0,0} + \sum_{(k,l) \in E} \left[A_{k,l} \cos\left[2\pi \left(\frac{k}{A}x + \frac{l}{B}y\right)\right] + B_{k,l} \sin\left[2\pi \left(\frac{k}{A}x + \frac{l}{B}y\right)\right]\right],$$

where the set $E \subset \mathbb{Z}^2$ is given by $E = (\mathbb{N} \times \mathbb{Z}) \cup (\{0\} \times \mathbb{N})$.

Putting these Fourier series together gives

$$\boldsymbol{v}(\boldsymbol{s}) = \begin{bmatrix} A^{(1)}_{0,0} \\ A^{(2)}_{0,0} \end{bmatrix} + \sum_{(k,l)\in E} \begin{bmatrix} A^{(1)}_{k,l} \\ A^{(2)}_{k,l} \end{bmatrix} \cos\left[2\pi\left(\frac{k}{A}x + \frac{l}{B}y\right)\right] + \sum_{(k,l)\in E} \begin{bmatrix} B^{(1)}_{k,l} \\ B^{(2)}_{k,l} \end{bmatrix} \sin\left[2\pi\left(\frac{k}{A}x + \frac{l}{B}y\right)\right],$$

where $A^{(1)}_{k,l}$ and $B^{(1)}_{k,l}$ are the coefficients for the first component of $\boldsymbol{v}$ and $A^{(2)}_{k,l}$ and $B^{(2)}_{k,l}$ are the coefficients of the second component. This gives 2 coefficients when only the zero-frequency is included, then 18 parameters when the $(0,1)$, $(1,-1)$, $(1,0)$ and $(1,1)$ frequencies are included. When the number of frequencies used in each direction doubles, the number of required parameters quadruples.

4.2 Inference on simulated data

In the supplementary material, we consider data generated from known sets of parameters for models of the type

$$u(\boldsymbol{s}) - \nabla \cdot \mathbf{H}(\boldsymbol{s})\nabla u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}),$$

where $\mathcal{W}$ is a standard Gaussian white noise process and $\mathbf{H}(\cdot)$ is a spatially varying 2×2 matrix, with periodic boundary conditions on a rectangular domain. The matrix is parameterized as $\mathbf{H}(\boldsymbol{s}) = \gamma \boldsymbol{I} + \boldsymbol{v}(\boldsymbol{s})\boldsymbol{v}(\boldsymbol{s})^{\mathrm{T}}$. The results illustrate the ability to estimate the vector field controlling the anisotropy for four test cases.

These examples focus on simple cases where specific issues can be highlighted. The inherent challenges in estimating a spatially varying direction and strength are equally important in the more general setting where also κ and the baseline effect γ are allowed to vary. The estimation of the vector field presents an important component that must be dealt with in any inference strategy for the more general case.

5 Extensions

To make the model applicable to datasets it is necessary to also make the parameters κ and γ spatially varying functions. This results in some control also over the marginal variance and the strength of the local baseline component of the anisotropy at each location. A varying κ is discussed briefly in Section 3.2 in Lindgren, Rue, and Lindström (2011).

This comes at the cost of two more functions that must be inferred together with the vector field $\boldsymbol{v}$ that, in turn, means two more functions need to be expanded into bases. This could be done with a Fourier basis, but any basis which respects the boundary condition could in principle be used. The amount of freedom available by having four spatially varying functions comes at a price, and it would be necessary to introduce some apriori restrictions on their behaviour.

In the supplementary material we demonstrate the challenge with the non-identifiability of the sign of the vector field. It would be possible to make the situation less problematic by enforcing more structure in the estimated vector field. For example, through spline penalties which adds a preference for components without abrupt changes. Such apriori restrictions make sense both from a modelling perspective, in the sense that the properties should not change to quickly, and from a computational perspective, in the sense that it is desirable to avoid situations as the one encountered in the previous section where the direction of the vector field flips.

The full model could be used in an application through a three-step approach. First, choose an appropriate basis to use for each function and select an appropriate prior. This means deciding how many basis elements one is willing to use from a computational point of view, and how strong the apriori penalties need to be. Second, find the maximum aposteriori estimate of the functions κ, γ, v_1, and v_2. Third, assume the maximum aposteriori estimates are the true functions and calculate the predicted values and prediction variances. Full details of such an approach are beyond the current scope. This is being studied in current work on an application to annual rainfall data in the conterminous US (Fuglstad, Simpson, Lindgren, and Rue (2013)).

Another way forward deals with the interactions of the functions κ, γ,

v_1, and v_2. The functions interact in difficult ways to control marginal variance and to control anisotropy. As seen in Example 3.2 the vector field that controls the anisotropic behaviour is also linked to the marginal variances of the field. It would be desirable to try to separate the functions that are allowed to affect the marginal variances and the functions that are allowed to affect the correlation structure. This may present a useful feature in applications, both for interpretability and for constructing priors.

One promising way to reduce interaction is to extend ideas presented in Section 3.4 of Lindgren, Rue, and Lindström (2011), linking the use of an anisotropic Laplacian to the deformation method of Sampson and Guttorp (1992). The link is too restrictive, but the last comments about the connection to metric tensors leads to a useful way to rewrite the SPDE in (2.3). This is work in progress and involves interpreting the simple SPDE

$$[1 - \Delta]u = \mathcal{W}$$

as an SPDE on a Riemannian manifold with an inverse metric tensor defined through the strength of dependence in different directions, in a similar way as the spatially varying matrix $\mathbf{H}$. This leads to a slightly different SPDE where a separate function, that does not affect correlation structure, can be used to control marginal standard deviations. However, the separation is not perfect since the varying metric tensor gives a curved space and thus affects the marginal variances of the solution of the above SPDE, though the effect of the metric tensor on marginal standard deviations appears small.

Another issue which has not been addressed is how to define relevant boundary conditions. For rectangular domains, periodic boundary conditions are simple to implement, but a naive use of such conditions is typically inappropriate in practical applications due to the resulting spurious dependence between physically distant locations. This problem can be partly rectified by embedding the region of interest into a larger covering domain. It is also possible to apply Neumann-type boundary conditions similar to the ones used by Lindgren, Rue, and Lindström (2011). These are easier to adapt to more general domains, but they still require a domain extension in order to remove the influence of the boundary condition on the likelihood. A more theoretically appealing, and computationally

potentially less expensive, solution is to directly define the behaviour of the field along the boundary so that the models would contain stationary fields as a neutral case. Work is underway to design stochastic boundary conditions to accomplish this, and some of the solutions show potential for extension to non-stationary models.

6 Discussion

The paper explores different aspects of a new class of non-stationary GRFs based on local anisotropy. The benefit of the formulation presented is that it allows for flexible models with few requirements on the parameters. Since the GRF is based on an SPDE, there is no need to worry about how to change the discretized model in a consistent manner when the grid is refined. This is one of the more attractive features of the SPDE-based modelling.

The focus of the examples has been the matrix $\mathbf{H}$ introduced in the Laplace-operator. The examples show that a variety of different effects can be achieved by using different types of spatially varying matrices. As shown in Section 3, anisotropic fields have anisotropic Matérn-like covariance functions, through stretching and rotating the domain, and can be controlled by four parameters. It is possible to control the marginal variances, the principal directions, and the range in each of the principal directions. A spatially varying $\mathbf{H}$ gives non-stationary random fields. And by using a vector field to specify the strength and direction of extra spatial dependence in each location, there is a clear connection between the vector field and the resulting covariance structure. The covariance structure can be partially visualized from the vector field.

There are many avenues that are not explored here. The chief motivation is to explore a class of models for what can be achieved, and for the associated challenges of inference with the models. We show that a vector field constitutes a useful way to control local anisotropy in the SPDE-model of Lindgren, Rue, and Lindström (2011). What remains for a fully flexible spatial model is to allow κ and γ to be spatially varying functions, this is a simpler task than the anisotropy component since they do not require vector fields. For this more complex model there are four spatially varying functions to estimate and an expansion of each of these

functions into a basis leads to many parameters. It remains to investigate appropriate choices of priors for use in applications. This question is connected with the discussion in Section 5 on other constructions of the model that separate the functions allowed to affect marginal variances and the functions allowed to affect correlation structure.

Acknowledgements

The authors are grateful to the Guest Editors of the special issue, and to the referees for their helpful comments which improved the manuscript.

References

Adler, R.J. and Taylor, J.E. (2007). *Random fields and geometry*. Springer Verlag.

Anderes, Ethan B and Stein, Michael L (2008). Estimating deformations of isotropic Gaussian random fields on the plane. *The Annals of Statistics* **31**, 719–741.

Banerjee, Sudipto, Gelfand, Alan E., Finley, Andrew O., and Sang, Huiyan (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B* **70**, 825–848.

Besag, Julian (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B* **2**, 192–236.

Bolin, D. and Lindgren, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *The Annals of Applied Statistics* **5**, 523–550.

Bolin, David (2013). Spatial Matérn Fields Driven by Non-Gaussian Noise. *Scandinavian Journal of Statistics* In press.

Cressie, Noel and Johannesson, Gardar (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B* **70**, 209–226.

Dahlhaus, R and Künsch, H (1987). Edge effects and efficient parameter estimation for stationary random fields. *Biometrika* **74**, 877–882.

Damian, Doris, Sampson, Paul D, and Guttorp, Peter (2001). Bayesian estimation of semi-parametric non-stationary spatial covariance structures. *Environmetrics* **12**, 161–178.

Damian, Doris, Sampson, Paul D., and Guttorp, Peter (2003). Variance model-

ing for nonstationary spatial processes with temporal replications. *Journal of Geophysical Research: Atmospheres* **108**.

Fuentes, Montserrat (2001). A high frequency kriging approach for non-stationary environmental processes. *Environmetrics* **12**, 469–483.

— (2002a). Interpolation of nonstationary air pollution processes: a spatial spectral approach. *Statistical Modelling* **2**, 281–298.

— (2002b). Spectral methods for nonstationary spatial processes. *Biometrika* **89**, 197–210.

— (2007). Approximate Likelihood for Large Irregularly Spaced Spatial Data. *Journal of the American Statistical Association* **102**, 321–331.

Fuglstad, Geir-Arne, Simpson, Daniel, Lindgren, Finn, and Rue, Håvard (2013). Non-stationary Spatial Modelling with Applications to Spatial Prediction of Precipitation. arXiv:1306.0408. In preperation.

Furrer, Reinhard, Genton, Marc G, and Nychka, Douglas (2006). Covariance Tapering for Interpolation of Large Spatial Datasets. *Journal of Computational and Graphical Statistics* **15**, 502–523.

Haas, Timothy C. (1990a). Kriging and automated variogram modeling within a moving window. *Atmospheric Environment. Part A. General Topics* **24**, 1759 – 1769.

— (1990b). Lognormal and Moving Window Methods of Estimating Acid Deposition. *Journal of the American Statistical Association* **85**, 950–963.

Higdon, David (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics* **5**, 173–190.

Ingebrigtsen, Rikke, Lindgren, Finn, and Steinsland, Ingelin (2013). Spatial Models with Explanatory Variables in the Dependence Structure of Gaussian Random Fields based on Stochastic Partial Differential Equations. *Spatial Statistics*. In press.

Kim, Hyoung-Moon, Mallick, Bani K, and Holmes, C. C (2005). Analyzing Nonstationary Spatial Data Using Piecewise Gaussian Processes. *Journal of the American Statistical Association* **100**, 653–668.

Lindgren, Finn and Rue, Håvard (2008). On the Second-Order Random Walk Model for Irregular Locations. *Scandinavian journal of statistics* **35**, 691–700.

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

Nychka, Douglas, Wikle, Christopher, and Royle, J Andrew (2002). Multireso-

lution models for nonstationary spatial covariance functions. *Statistical Modelling* **2**, 315–331.

Paciorek, Christopher J. and Schervish, Mark J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics* **17**, 483–506.

Rue, Håvard and Held, Leonard (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Rue, Håvard, Martino, Sara, and Chopin, Nicolas (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B* **71**, 319–392.

Sampson, Paul D. and Guttorp, Peter (1992). Nonparametric Estimation of Nonstationary Spatial Covariance Structure. *Journal of the American Statistical Association* **87**, 108–119.

Schmidt, Alexandra M., Guttorp, Peter, and O'Hagan, Anthony (2011). Considering covariates in the covariance structure of spatial processes. *Environmetrics* **22**, 487–500.

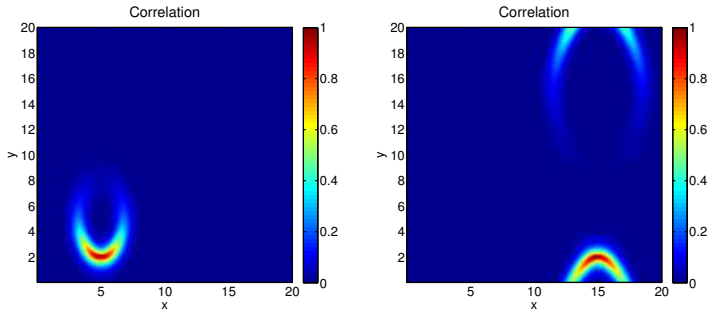
Schmidt, Alexandra M. and O'Hagan, Anthony (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B* **65**, 743–758.

Stein, Michael L., Chi, Zhiyi, and Welty, Leah J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B* **66**, 275–296.

Sun, Ying, Li, Bo, and Genton, Marc G. (2012). Geostatistics for Large Datasets. *Advances and Challenges in Space-time Modelling of Natural Events*, 55–77. Springer Berlin Heidelberg.

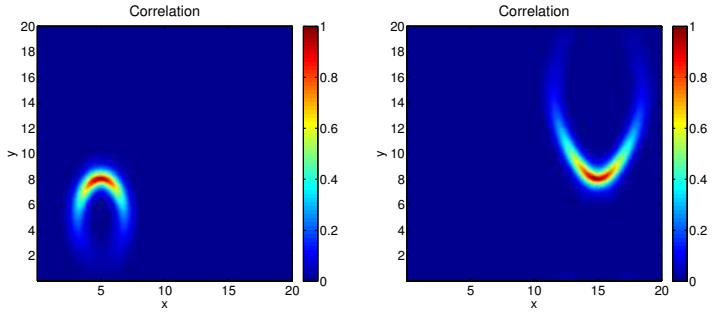
Whittle, P. (1954). On Stationary Processes in the Plane. *Biometrika* **41**, 434–449.

Department of Mathematical Sciences, NTNU, 7491 Trondheim, Norway
E-mail: fuglstad@math.ntnu.no
Department of Mathematical Sciences, University of Bath, United Kingdom
E-mail: f.lindgren@bath.ac.uk
Department of Mathematical Sciences, NTNU, Norway
E-mail: dp.simpson@gmail.com
Department of Mathematical Sciences, NTNU, Norway
E-mail: hrue@math.ntnu.no



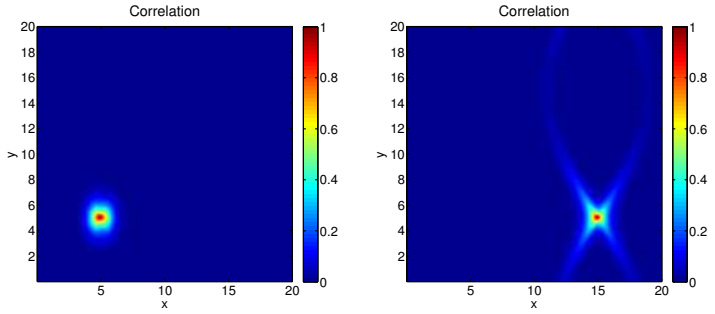
(a) Correlations with position $(4.95, 1.95)$.

(b) Correlations with position $(14.95, 2.05)$.



(c) Correlations with position $(4.95, 7.95)$.

(d) Correlations with position $(14.95, 7.95)$.



(e) Correlations with position $(4.95, 4.95)$.

(f) Correlations with position $(14.95, 4.95)$.

Figure 5: Correlations for different points with all other points for the solution of the SPDE in Example 3.2.

Supplementary material:
Exploring a New Class of Non-stationary Spatial Gaussian Random Fields with Varying Local Anisotropy

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

[1]Department of Mathematical Sciences, NTNU, Norway
[2]Department of Mathematical Sciences, University of Bath, United Kingdom

The supplementary material contains investigation of inference on simulated data, the derivation of the discretization of the SPDE used in the paper, and the derivation of the marginal variance of the stationary solution.

S1 Inference on simulated data

We consider data generated from a known set of parameters. The prior used is improper, uniform on $(0, \infty)$ for γ and uniform on $\mathbb{R}$ for the rest of the parameters in $\mathbf{H}$. We investigate whether it is possible to estimate the stationary model with exactly observed data and whether the approximate estimation scheme performs well.

Example S1.1. Use the SPDE

$$u(\boldsymbol{s}) - \nabla \cdot \mathbf{H} \nabla u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \qquad \boldsymbol{s} \in [0, 20] \times [0, 20], \tag{S1.1}$$

where $\mathcal{W}$ is a standard Gaussian white noise process and $\mathbf{H}$ is a 2×2 matrix, with periodic boundary conditions. Let $\mathbf{H} = 3\mathbf{I}_2 + 2\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}$, with

Table 1: Parameter estimates for Example S1.1.

Parameter	**True value**	**Estimate**	**Std.dev.**
γ	3	2.965	0.070
v_1	0.707	0.726	0.049
v_2	1.225	1.231	0.039

$\boldsymbol{v} = (1, \sqrt{3})/2$. Here $\mathbf{H}$ has an eigenvector $\boldsymbol{v}$ with eigenvalue 5 and an eigenvector orthogonal to $\boldsymbol{v}$ with eigenvalue 3. Construct the GMRF on a 100×100 grid.

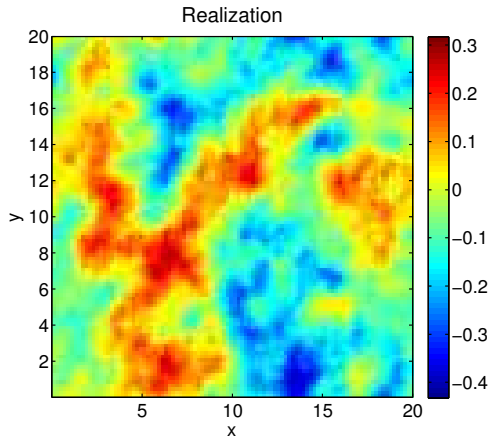
One observation of the solution is shown in Figure 1(a). Assume that $\mathbf{H}$ is constant, but that its value is not known. Then using the decomposition from the previous sections one can write

$$\mathbf{H} = \gamma \mathbf{I}_2 + \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \begin{bmatrix} v_1 & v_2 \end{bmatrix},$$

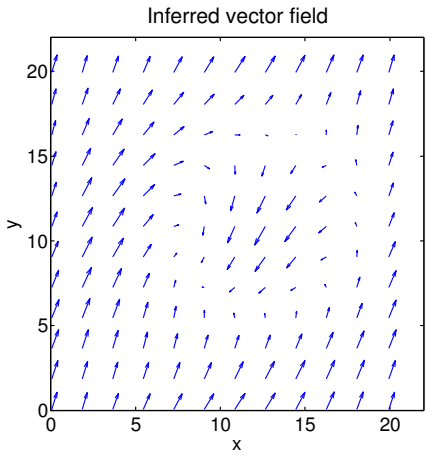
where γ, v_1 and v_2 are the parameters. Since the process is assumed to be exactly observed, we can use the distribution of $\boldsymbol{\theta}|\boldsymbol{u}$. This gives the posterior estimates shown in Table 1. From the table one can see that all the estimates are accurate to one digit, and within one standard deviation of the true value. This decomposition of $\mathbf{H}$ is invariant to the sign of $\boldsymbol{v}$, so there are two choices of parameters that mean the same.

The biases in the estimates were evaluated by generating 10000 datasets from the true model and estimating the parameters for each dataset. The estimated bias was less than or equal to 0.1% of the true value for each parameter. Additionally, the sample standard deviations based on the estimation of the parameters for each of the 10000 datasets were 0.070, 0.050, and 0.039 for γ, v_1, and v_2, respectively. Each one corresponds well to the corresponding approximate standard deviation, computed via the observed information matrix, that is shown in Table 1.

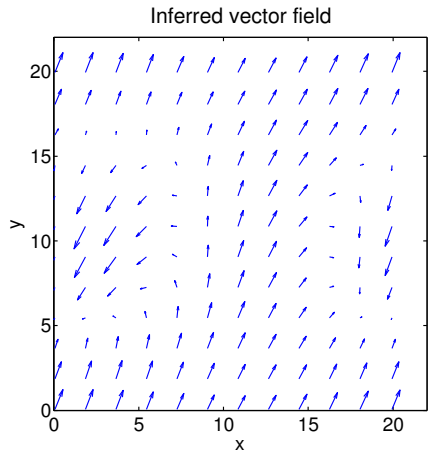
It is then possible to estimate the model, but this is under the assumption that it is known beforehand that the model is stationary. The estimation is repeated for the more complex model developed in the previous sections that allows for significant non-stationarity controlled through a vector field. The intention is to evaluate whether the more complex



(a) Example S1.1.



(b) Wrong maximum.



(c) Wrong maximum.

Figure 1: (a) One realization of the solution of the SPDE in Example S1.1. (b) and (c) show two local maxima found for the vector field in Example S1.2. The vector field in (b) has a lower value for the posterior distribution of the parameters than the vector field in (c) and both have lower value than the actual maximum.

model is able to detect that the true model is a stationary model, and if there are identifiability issues.

Example S1.2. Use the same SPDE and observation as in Example S1.1, but assume that it is not known that $\mathbf{H}$ is constant. Add the terms in the Fourier series corresponding to the next frequencies, $(k, l) = (0, 1)$, $(k, l) = (1, -1)$, $(k, l) = (1, 0)$ and $(k, l) = (1, 1)$. The observation is still assumed to be exact, but there are 16 additional parameters, 4 additional parameters for each frequency.

Two arbitrary starting positions are chosen for the optimization: $\gamma = 3.0$ and all other parameters at 0.1; $\gamma = 3.0$, $A^{(1)}_{0,0} = 0.1$, $A^{(2)}_{0,0} = 0.1$, and all other parameters equal to 0. For both starting points the optimization converges to non-global maximums. Parameter estimates and approximate standard deviations are not given, but Figures 1(b) and 1(c) show the two vector fields found.

A third optimization is done with starting values close to the correct parameter values. This gives a vector field close to the actual one, with estimates for γ, $A^{(1)}_{0,0}$, and $A^{(2)}_{0,0}$ that agree with the ones in Example S1.1 to two digits. The other frequencies all had coefficients close to zero, with the largest having an absolute value of 0.058.

The results illustrate a difficulty with estimation caused by the the inherent non-identifiability of the sign of the vector field. The true vector field is constant, and each estimated vector field has large parts which have the correct appearance if one only considers the lines defined by the arrows and not the direction. The positions where the vector field is wrong are smaller areas where the vector field flips its direction. It is difficult to reverse this flipping as it requires moving through states with smaller likelihood, thus creating undesirable local maximums. One approach to improving the situation would be to force an apriori preference for vector fields without abrupt changes, to introduce a prior that forces higher frequencies of the Fourier basis to be less desirable. This is an issue that needs to be addressed, and is briefly discussed in the main paper.

By starting close to the true value, one can do repeated simulations of datasets and prediction of parameters to evaluate how well the non-stationary model captures the fact that the true model is stationary and to see if there is any consistent bias. 1000 datasets were simulated and the estimation of the parameters was done for each dataset with a starting

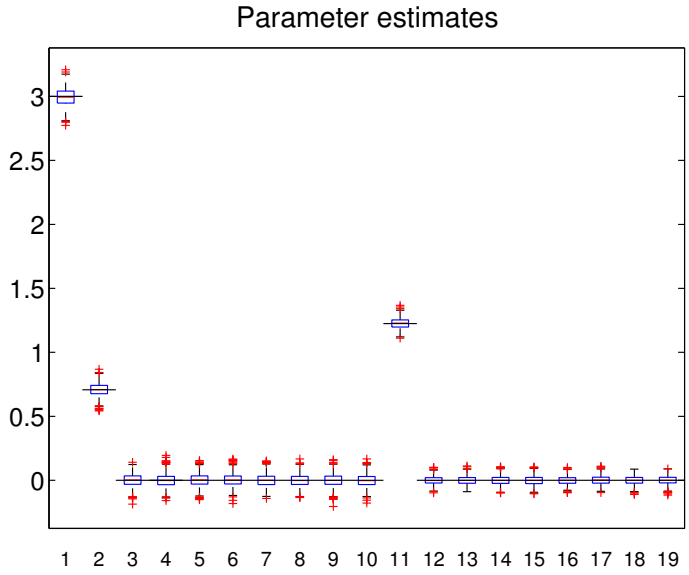


Figure 2: Boxplot of estimated parameters for 1000 simulated datasets in Example S1.2. Parameters 1, 2 and 11 corresponds to γ, v_1 and v_2, respectively. The horizontal lines through the boxes specify the true parameter values, which in most cases closely match the medians.

value close to the true value. This gives the result summarized in the boxplot in Figure 2. There does not appear to be any significant bias and the parameters that give non-stationarity are all close to zero.

There are issues in estimating the anisotropy in the non-stationary model due to the non-identifiability of the sign of the vector field, but if one avoids the local maximums the estimated model is close to the true stationary model. There is a significant increase in computation time when increasing the parameter space from 3 to 19 parameters; the time required increases by a factor of approximately 10.

In situations where there is a physical explanation of the additional dependence in one direction, it would be desirable to do a simpler model with one parameter for the baseline isotropic effect and one parameter specifying the degree of anisotropy caused by a pre-defined vector field.

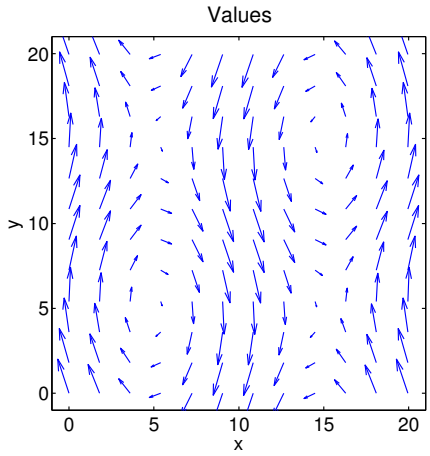


Figure 3: Vector field used in Example S1.3.

Table 2: Posterior inference on parameters in Example S1.3.

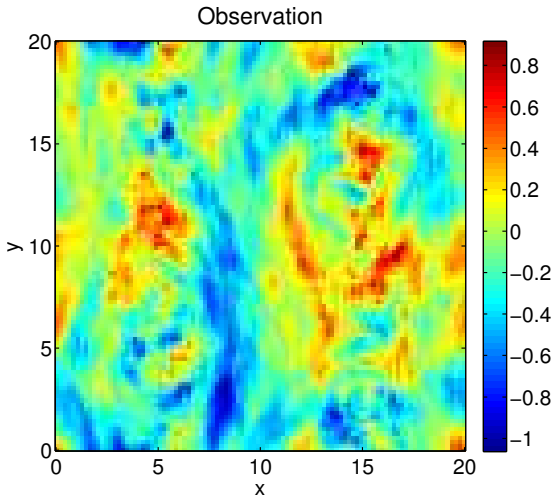
Parameter	**True value**	**Estimate**	**Std.dev.**
γ	0.5	0.5012	0.0081
β	5	5.014	0.084

Example S1.3. Use a 100×100 grid of $[0, 20]^2$ and periodic boundary conditions for the SPDE in (S1.1). Let κ^2 be 1 and let $\mathbf{H}(\boldsymbol{s}) = \gamma \mathbf{I}_2 + \beta \boldsymbol{v}(\boldsymbol{s})\boldsymbol{v}(\boldsymbol{s})^{\mathrm{T}}$, where $\boldsymbol{v}$ is the vector field in Figure 3.

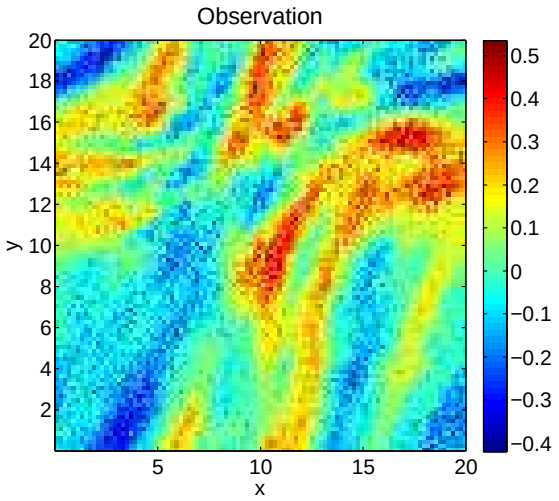
Figure 4(a) shows one observation of the solution with $\gamma = 0.5$ and $\beta = 5$. One expects that it is possible to make accurate estimates about γ and β as the situation is simpler than in the previous example.

The estimated parameters are shown in Table 2. From the table one can see that the estimates for γ and β are quite accurate, to 2 digits. As in the previous example, the bias is estimated to be less than 0.02% for each parameter, and the sample standard deviation from estimation over many datasets is 0.008 and 0.08 for γ and β, respectively.

Example S1.4. Use a 100×100 grid of $[0, 20]^2$ and periodic boundary conditions for the SPDE in (S1.1). Let κ^2 be 1 and let $\mathbf{H}(\boldsymbol{s}) = \mathbf{I}_2 +$



(a) Example S1.3



(b) Example S1.4

Figure 4: (a) An observation of the SPDE in (S1.1) on a 100×100 regular grid of $[0, 20]^2$ with periodic boundary conditions, $\kappa^2 = 1$ and $\mathbf{H}(\boldsymbol{s}) = 0.5\mathbf{I}_2 + 5\boldsymbol{v}(\boldsymbol{s})\boldsymbol{v}(\boldsymbol{s})^{\mathrm{T}}$, where $\boldsymbol{v}$ is the vector field in Figure 3. (b) An observation of the SPDE in Example S1.4 with i.i.d. Gaussian white noise with precision 400.

$\boldsymbol{v}(\boldsymbol{s})\boldsymbol{v}(\boldsymbol{s})^{\mathrm{T}}$, where $\boldsymbol{v}$ is the vector field

$$\boldsymbol{v}(x, y) = \begin{bmatrix} 2 + \cos\left(\frac{\pi}{10}x\right) \\ 3 + 2\sin\left(\frac{\pi}{10}y\right) + \sin\left(\frac{\pi}{10}(x+y)\right) \end{bmatrix}.$$

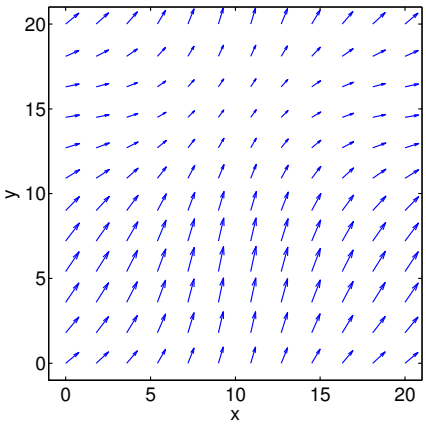
One observation with i.i.d. Gaussian noise with precision 400 is shown in Figure 4(b). Based on this realization it is desired to estimate the correct value of γ and the correct vector field $\boldsymbol{v}$ in the parametrization $\mathbf{H}(\boldsymbol{s}) = \gamma\mathbf{I}_2 + \boldsymbol{v}(\boldsymbol{s})\boldsymbol{v}(\boldsymbol{s})^{\mathrm{T}}$. First use only the frequencies $(0,0)$, $(0,1)$, and $(1,0)$. This gives the estimated vector field shown in Figure 5(a). Then use the frequencies $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$. This gives the estimated vector field shown in Figure 5(b). The true vector field is shown in Figure 5(c).

The estimated vector fields are quite similar to the true vector field, and the γ parameter estimate was 1.14 in the first case and 1.09 in the latter case. There is a clear bias in the estimate of γ, to be expected as there is a need to compensate for the lacking frequencies. All parameter values were estimated, but are not shown. For the first case many parameters are more than two standard deviations from their correct values and, in the second case, this only happens for one parameter. If the difference between the true $\mathbf{H}$ and the estimated $\hat{\mathbf{H}}$ is calculated through

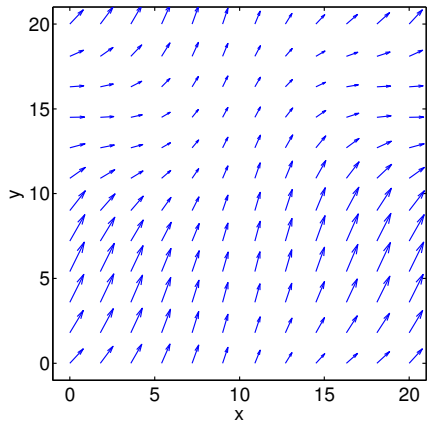
$$\frac{1}{100}\sqrt{\sum_{i=1}^{100}\sum_{j=1}^{100}\left|\left|\mathbf{H}(\boldsymbol{s}_{i,j}) - \hat{\mathbf{H}}(\boldsymbol{s}_{i,j})\right|\right|_2^2},$$

where $\boldsymbol{s}_{i,j}$ are the centres of the cells in the grid and $||\cdot||_2$ denotes the 2-norm, the first case gives 7.9 and the second case gives 1.5.

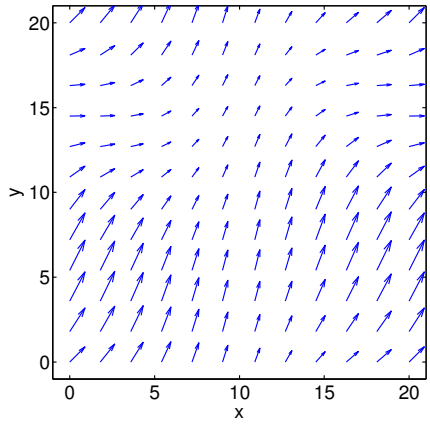
These examples focus on simple cases where specific issues can be highlighted. The inherent challenges in estimating a spatially varying direction and strength are equally important in the more general setting where also κ and the baseline effect γ is allowed to vary. The estimation of the vector field presents an important component that must be dealt with in any inference strategy for the more general case.



(a) $(0,0)$, $(0,1)$ and $(1,0)$ frequencies



(b) $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$ frequencies



(c) The true vector field

Figure 5: True vector field and inferred vector fields in Example S1.4. Each of the vector fields is scaled with a factor 0.3.

S2 Derivation of precision matrix

S2.1 Formal equation

The SPDE is

$$(\kappa^2(\boldsymbol{s}) - \nabla \cdot \mathbf{H}(\boldsymbol{s}))\nabla u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \qquad \boldsymbol{s} \in [0, A] \times [0, B], \tag{S2.1}$$

where A and B are strictly positive constants, κ^2 is a scalar function, $\mathbf{H}$ is a 2×2 matrix-valued function, $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$, and $\mathcal{W}$ is a standard Gaussian white noise process. Here κ^2 is assumed to be a continuous, strictly positive function and $\mathbf{H}$ is assumed to be a continuously differentiable function which gives a positive definite matrix $\mathbf{H}(\boldsymbol{s})$ for each $\boldsymbol{s} \in [0, A] \times [0, B]$.

Periodic boundary conditions are used so that opposite sides of the rectangle $[0, A] \times [0, B]$ are identified. Thus values of κ^2 must agree on opposite edges, and the values of $\mathbf{H}$ and its first order derivatives must agree on opposite edges. Periodic boundary conditions are not essential to the methodology presented, but we avoid the issue of appropriate boundary conditions for now.

S2.2 Finite volume methods

In the discretization of the SPDE in (S2.1), a finite volume method is employed. Finite volume methods are useful for creating discretizations of conservation laws of the form

$$\nabla \cdot \boldsymbol{F}(\boldsymbol{x}, t) = f(\boldsymbol{x}, t),$$

where $\nabla\cdot$ is the spatial divergence operator. This equation relates the spatial divergence of the flux $\boldsymbol{F}$ and the sink-/source-term f. The main tool here is the use of the divergence theorem

$$\int_E \nabla \cdot \boldsymbol{F} \, \mathrm{d}V = \oint_{\partial E} \boldsymbol{F} \cdot \boldsymbol{n} \, \mathrm{d}\sigma, \tag{S2.2}$$

where $\boldsymbol{n}$ is the outer normal vector of the surface ∂E relative to E.

The main idea is to divide the domain of the SPDE in (S2.1) into smaller parts and consider the resulting "flow" between the different parts. A lengthy treatment of finite volume methods is not given, but a comprehensive treatment of the method for deterministic differential equations can be found in Eymard, Gallouët, and Herbin (2000).

S2.3 Derivation

To keep the calculations simple the domain is divided into a regular grid of rectangular cells. Use M cells in the x-direction and N cells in the y-direction. Then for each cell the sides parallel to the x-axis have length $h_x = A/M$ and the sides parallel to the y-axis have length $h_y = B/N$. Number the cells by (i, j), where i is the column of the cell (along the x-axis) and j is the row of the cell (along the y-axis). If the lowest row is 0 and the leftmost column is 0, then cell (i, j) is

$$E_{i,j} = [ih_x, (i+1)h_x] \times [jh_y, (j+1)h_y],$$

and the set of cells, $\mathcal{I}$, is

$$\mathcal{I} = \{E_{i,j} : i = 0, 1, \ldots, M-1, j = 0, 1, \ldots, N-1\}.$$

Figure 6 shows an illustration of the discretization of $[0, A] \times [0, B]$ into the cells $\mathcal{I}$.

Each cell has four faces, two parallel to the x-axis (top and bottom) and two parallel to the y-axis (left and right). Let the right face, top face, left face and bottom face of cell $E_{i,j}$ be denoted $\sigma^{\mathrm{R}}_{i,j}$, $\sigma^{\mathrm{T}}_{i,j}$, $\sigma^{\mathrm{L}}_{i,j}$ and $\sigma^{\mathrm{B}}_{i,j}$, respectively. Additionally, denote by $\sigma(E_{i,j})$ the set of faces of cell $E_{i,j}$.

For each cell $E_{i,j}$, $\boldsymbol{s}_{i,j}$ gives the centroid of the cell, and $\boldsymbol{s}_{i+1/2,j}$, $\boldsymbol{s}_{i,j+1/2}$, $\boldsymbol{s}_{i-1/2,j}$, and $\boldsymbol{s}_{i,j-1/2}$ give the centres of the faces of the cell. Due to the periodic boundary conditions, the i-index and j-index in $\boldsymbol{s}_{i,j}$ are modulo M and modulo N, respectively. Figure 7 shows one cell $E_{i,j}$ with the centroid and the faces marked on the figure. Further, let $u_{i,j} = u(\boldsymbol{s}_{i,j})$ for each cell and denote the area of $E_{i,j}$ by $V_{i,j}$. Since the grid is regular, all $V_{i,j}$ are equal to $V = h_x h_y$.

To derive the finite volume scheme, begin by integrating (S2.1) over a cell, $E_{i,j}$. This gives

$$\int_{E_{i,j}} \kappa^2(\boldsymbol{s}) u(\boldsymbol{s}) \, \mathrm{d}\boldsymbol{s} - \int_{E_{i,j}} \nabla \cdot \mathbf{H}(\boldsymbol{s}) \nabla u(\boldsymbol{s}) \, \mathrm{d}\boldsymbol{s} = \int_{E_{i,j}} \mathcal{W}(\boldsymbol{s}) \, \mathrm{d}\boldsymbol{s}, \qquad \text{(S2.3)}$$

where $\mathrm{d}\boldsymbol{s}$ is an area element. The integral on the right hand side is distributed as a Gaussian variable with mean 0 and variance V for each (i, j) (Adler and Taylor (2007, pp. 24–25)). Further, the integral on the

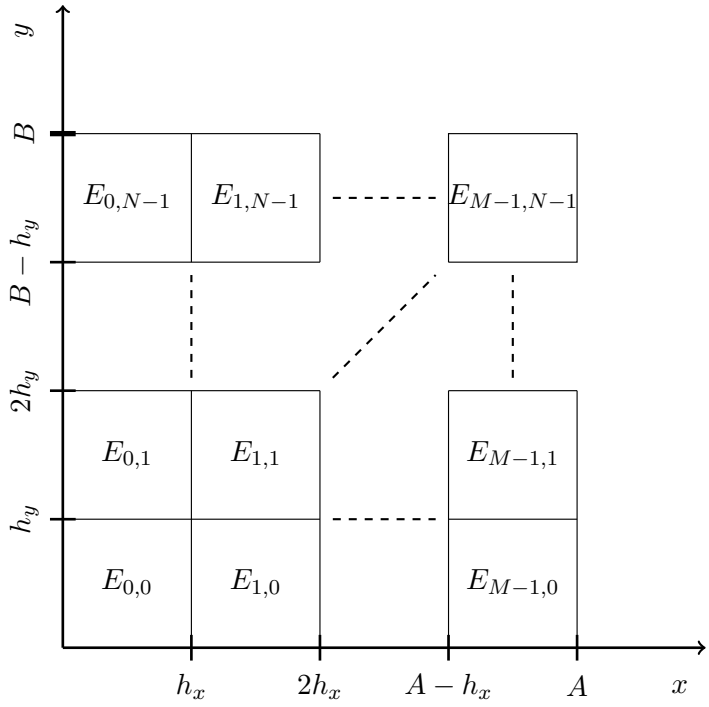


Figure 6: Illustration of the division of $[0, A] \times [0, B]$ into a regular $M \times N$ grid of rectangular cells.

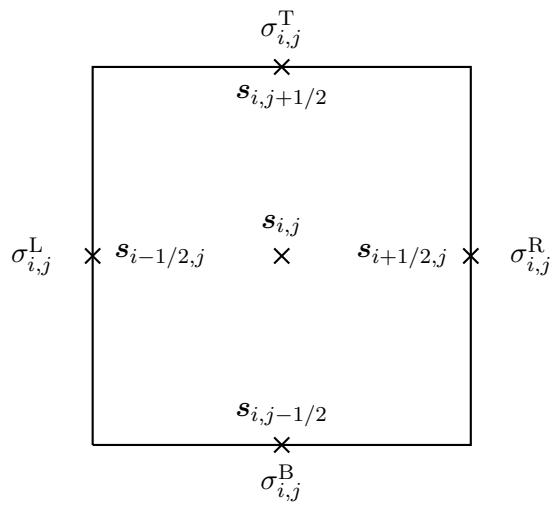


Figure 7: One cell, $E_{i,j}$, of the discretization with faces $\sigma^{\mathrm{R}}_{i,j}$, $\sigma^{\mathrm{T}}_{i,j}$, $\sigma^{\mathrm{L}}_{i,j}$ and $\sigma^{\mathrm{B}}_{i,j}$, centroid $\boldsymbol{s}_{i,j}$ and centres of the faces $\boldsymbol{s}_{i-1/2,j}$, $\boldsymbol{s}_{i,j-1/2}$, $\boldsymbol{s}_{i+1/2,j}$ and $\boldsymbol{s}_{i,j+1/2}$.

right hand side is independent for different cells, because two different cells can at most share a common face. Thus (S2.3) can be written as

$$\int_{E_{i,j}} \kappa^2(\boldsymbol{s})u(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s} - \int_{E_{i,j}} \nabla\cdot\mathbf{H}(\boldsymbol{s})\nabla u(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s} = \sqrt{V}z_{i,j},$$

where $z_{i,j}$ is a standard Gaussian variable for each (i,j) and the Gaussian variables are independent.

By the divergence theorem, (S2.2), the second integral on the left hand side can be written as an integral over the boundary of the cell, so

$$\int_{E_{i,j}} \kappa^2(\boldsymbol{s})u(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s} - \oint_{\partial E_{i,j}} (\mathbf{H}(\boldsymbol{s})\nabla u(\boldsymbol{s}))^{\mathrm{T}}\boldsymbol{n}(\boldsymbol{s})\,\mathrm{d}\sigma = \sqrt{V}z_{i,j}, \qquad \text{(S2.4)}$$

where $\boldsymbol{n}$ is the exterior normal vector of $\partial E_{i,j}$ with respect to $E_{i,j}$ and $\mathrm{d}\sigma$ is a line element. It is useful to divide the integral over the boundary in Equation (S2.4) into integrals over each face,

$$\int_{E_{i,j}} \kappa^2(\boldsymbol{s})u(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s} - \left(W^R_{i,j} + W^T_{i,j} + W^L_{i,j} + W^B_{i,j}\right) = \sqrt{V}z_{i,j}, \qquad \text{(S2.5)}$$

where $W_{i,j}^{\text{dir}} = \int_{\sigma_{i,j}^{\text{dir}}} (\mathbf{H}(\boldsymbol{s})\nabla u(\boldsymbol{s}))^{\text{T}} \boldsymbol{n}(\boldsymbol{s}) \, \text{d}\sigma$.

The first integral on the left hand side of (S2.5) is approximated by

$$\int_{E_{i,j}} \kappa^2(\boldsymbol{s}) u(\boldsymbol{s}) \, \text{d}\boldsymbol{s} = V \kappa_{i,j}^2 u(\boldsymbol{s}_{i,j}) = V \kappa_{i,j}^2 u_{i,j}, \tag{S2.6}$$

where $\kappa_{i,j}^2 = \frac{1}{V} \int_{E_{i,j}} \kappa^2(\boldsymbol{s}) \, \text{d}\boldsymbol{s}$. The function κ^2 is assumed to be continuous and $\kappa_{i,j}^2$ is approximated by $\kappa^2(\boldsymbol{s}_{i,j})$.

The second part of (S2.5) requires the approximation of the surface integral over each face of a given cell. The values of $\mathbf{H}$ are in general not diagonal, so it is necessary to estimate both components of the gradient on each face of the cell. For simplicity, it is assumed that the gradient is constant on each face and that it is identically equal to the value at the centre of the face. On a face parallel to the y-axis the estimate of the partial derivative with respect to x is simple since the centroid of each of the cells which share the face have the same y-coordinate. The problem is the estimate of the partial derivative with respect to y. The reverse is true for the top and bottom face of the cell.

It is important to use a scheme which gives the same estimate of the gradient for a given face no matter which of the two neighbouring cells are chosen. For $\sigma_{i,j}^{\text{R}}$, the approximation used is

$$\frac{\partial}{\partial y} u(\boldsymbol{s}_{i+1/2,j}) \approx \frac{1}{h_y} (u(\boldsymbol{s}_{i+1/2,j+1/2}) - u(\boldsymbol{s}_{i+1/2,j-1/2})).$$

where the values of u at $\boldsymbol{s}_{i+1/2,j+1/2}$ and $\boldsymbol{s}_{i+1/2,j-1/2}$ are linearly interpolated from the values at the four closest cells. More precisely, because of the regularity of the grid the mean of the four closest cells are used, giving

$$\frac{\partial}{\partial y} u(\boldsymbol{s}_{i+1/2,j}) \approx \frac{1}{4h_y} (u_{i+1,j+1} + u_{i,j+1} - u_{i,j-1} - u_{i+1,j-1}). \tag{S2.7}$$

This formula can be used for the partial derivative with respect to y on any face parallel to the y-axis by suitably changing the i and j indices. The partial derivative with respect to x on a face parallel to the y-axis can be approximated directly by

$$\frac{\partial}{\partial x} u(\boldsymbol{s}_{i+1/2,j}) \approx \frac{1}{h_x} (u_{i+1,j} - u_{i,j}). \tag{S2.8}$$

Table 3: Finite difference schemes for the partial derivative with respect to x and y at the different faces of cell $E_{i,j}$.

Face	$\frac{\partial}{\partial x}u(s)$	$\frac{\partial}{\partial y}u(s)$
$\sigma_{i,j}^{\mathrm{R}}$	$\frac{u_{i+1,j}-u_{i,j}}{h_x}$	$\frac{u_{i,j+1}+u_{i+1,j+1}-u_{i,j-1}-u_{i+1,j-1}}{4h_y}$
$\sigma_{i,j}^{\mathrm{T}}$	$\frac{u_{i+1,j}+u_{i+1,j+1}-u_{i-1,j}-u_{i-1,j+1}}{4h_x}$	$\frac{u_{i,j+1}-u_{i,j}}{h_y}$
$\sigma_{i,j}^{\mathrm{L}}$	$\frac{u_{i,j}-u_{i-1,j}}{h_x}$	$\frac{u_{i-1,j+1}+u_{i,j+1}-u_{i-1,j-1}-u_{i,j-1}}{4h_y}$
$\sigma_{i,j}^{\mathrm{B}}$	$\frac{u_{i+1,j}+u_{i+1,j-1}-u_{i-1,j-1}-u_{i-1,j}}{4h_x}$	$\frac{u_{i,j}-u_{i,j-1}}{h_y}$

In more or less the same way the two components of the gradient on the top face of cell $E_{i,j}$ can be approximated by

$$\frac{\partial}{\partial x}u(\boldsymbol{s}_{i,j+1/2}) \approx \frac{1}{4h_x}(u_{i+1,j+1} + u_{i+1,j} - u_{i-1,j} - u_{i-1,j+1}),$$

$$\frac{\partial}{\partial y}u(\boldsymbol{s}_{i,j+1/2}) \approx \frac{1}{h_y}(u_{i,j+1} - u_{i,j}).$$

These approximations can be used on any side parallel to the x-axis by changing the indices appropriately.

The approximations for the partial derivatives on each face are collected in Table 3. Using these, one can find the approximations needed for the second part of (S2.5). It is helpful to write

$$W_{i,j}^{\mathrm{dir}} = \int_{\sigma_{i,j}^{\mathrm{dir}}} (\mathbf{H}(\boldsymbol{s})\nabla u(\boldsymbol{s}))^{\mathrm{T}}\boldsymbol{n}(\boldsymbol{s})\,\mathrm{d}\sigma = \int_{\sigma_{i,j}^{\mathrm{dir}}} (\nabla u(\boldsymbol{s}))^{\mathrm{T}}(\mathbf{H}(\boldsymbol{s})\boldsymbol{n}(\boldsymbol{s}))\,\mathrm{d}\sigma,$$

where the symmetry of $\mathbf{H}$ is used to avoid transposing the matrix. Assuming that the gradient is identically equal to the value at the centre of the face, one finds

$$W_{i,j}^{\mathrm{dir}} \approx \left(\nabla u(\boldsymbol{c}_{i,j}^{\mathrm{dir}})\right)^{\mathrm{T}} \int_{\sigma_{i,j}^{\mathrm{dir}}} \mathbf{H}(\boldsymbol{s})\boldsymbol{n}(\boldsymbol{s})\,\mathrm{d}\sigma,$$

where $\boldsymbol{c}_{i,j}^{\mathrm{dir}}$ is the centre of face $\sigma_{i,j}^{\mathrm{dir}}$.

Since the cells form a regular grid, $\boldsymbol{n}$ is constant on each face. If $\mathbf{H}$ be approximated by its value at the centre of the face, then

$$W_{i,j}^{\text{dir}} \approx m(\sigma_{i,j}^{\text{dir}}) \left(\nabla u(\boldsymbol{c}_{i,j}^{\text{dir}})\right)^{\text{T}} \left(\mathbf{H}(\boldsymbol{c}_{i,j}^{\text{dir}})\boldsymbol{n}(\boldsymbol{c}_{i,j}^{\text{dir}})\right), \tag{S2.9}$$

where $m(\sigma_{i,j}^{\text{dir}})$ is the length of the face. Note that the length of the face is either h_x or h_y and that the normal vector is parallel to the x-axis or the y-axis.

If

$$\mathbf{H}(\boldsymbol{s}) = \begin{bmatrix} H^{11}(\boldsymbol{s}) & H^{12}(\boldsymbol{s}) \\ H^{21}(\boldsymbol{s}) & H^{22}(\boldsymbol{s}) \end{bmatrix},$$

then using Table 3 one finds the approximations

$$\begin{aligned}
\hat{W}_{i,j}^{\text{R}} = & \\
& h_y \left[H^{11}(\boldsymbol{s}_{i+1/2,j}) \frac{u_{i+1,j} - u_{i,j}}{h_x} \right] + \\
& h_y \left[H^{21}(\boldsymbol{s}_{i+1/2,j}) \frac{u_{i,j+1} + u_{i+1,j+1} - u_{i,j-1} - u_{i+1,j-1}}{4h_y} \right],
\end{aligned}$$

$$\begin{aligned}
\hat{W}_{i,j}^{\text{T}} = & \\
& h_x \left[H^{12}(\boldsymbol{s}_{i,j+1/2}) \frac{u_{i+1,j+1} + u_{i+1,j} - u_{i-1,j+1} - u_{i-1,j}}{4h_x} \right] + \\
& h_x \left[H^{22}(\boldsymbol{s}_{i,j+1/2}) \frac{u_{i,j+1} - u_{i,j}}{h_y} \right],
\end{aligned}$$

$$\begin{aligned}
\hat{W}_{i,j}^{\text{L}} = & \\
& h_y \left[H^{11}(\boldsymbol{s}_{i-1/2,j}) \frac{u_{i-1,j} - u_{i,j}}{h_x} \right] + \\
& h_y \left[H^{21}(\boldsymbol{s}_{i-1/2,j}) \frac{u_{i,j-1} + u_{i-1,j-1} - u_{i-1,j+1} - u_{i,j+1}}{4h_y} \right],
\end{aligned}$$

$$\begin{aligned}
\hat{W}_{i,j}^{\text{B}} = & \\
& h_x \left[H^{12}(\boldsymbol{s}_{i,j-1/2}) \frac{u_{i-1,j} + u_{i-1,j-1} - u_{i+1,j} - u_{i+1,j-1}}{4h_x} \right] + \\
& h_x \left[H^{22}(\boldsymbol{s}_{i,j-1/2}) \frac{u_{i,j-1} - u_{i,j}}{h_y} \right].
\end{aligned}$$

These approximations can be combined with the approximations in (S2.6) and inserted into (S2.5) to give

$$V\kappa_{i,j}^2 u_{i,j} - \left(\hat{W}_{i,j}^{\mathrm{R}} + \hat{W}_{i,j}^{\mathrm{T}} + \hat{W}_{i,j}^{\mathrm{L}} + \hat{W}_{i,j}^{\mathrm{B}}\right) = \sqrt{V} z_{i,j}.$$

Stacking the variables $u_{i,j}$ row-wise in a vector $\boldsymbol{u}$ gives the linear system of equations,

$$\mathbf{D}_V \mathbf{D}_{\kappa^2} \boldsymbol{u} - \mathbf{A}_{\mathbf{H}} \boldsymbol{u} = \mathbf{D}_V^{1/2} \boldsymbol{z}, \tag{S2.10}$$

where $\mathbf{D}_V = V\mathbf{I}_{MN}$, $\mathbf{D}_{\kappa^2} = \mathrm{diag}(\kappa_{0,0}^2, \ldots, \kappa_{M-1,0}^2, \kappa_{0,1}^2, \ldots, \kappa_{M-1,N-1}^2)$, $\boldsymbol{z} \sim \mathcal{N}_{MN}(\mathbf{0}, \mathbf{I}_{MN})$; $\mathbf{A}_{\mathbf{H}}$ is considered more closely in what follows.

The construction of $\mathbf{A}_{\mathbf{H}}$, which depends on the function $\mathbf{H}$, requires only that one write out the sum

$$\hat{W}_{i,j}^{\mathrm{R}} + \hat{W}_{i,j}^{\mathrm{T}} + \hat{W}_{i,j}^{\mathrm{L}} + \hat{W}_{i,j}^{\mathrm{B}}$$

and collects the coefficients of the different $u_{a,b}$ terms. This is not difficult, but requires many lines of equations. Therefore, only the resulting coefficients are given. Fix (i, j) and consider the equation for cell $E_{i,j}$. For convenience, let i_p and i_n be the column left and right of the current column, respectively and let j_n and j_p be the row above and below the current row, respectively. These rows and columns are 0-indexed and due to the periodic boundary conditions one has, for example, that column 0 is to the right of column $M - 1$. Further, number the rows and columns of the matrix $\mathbf{A}_{\mathbf{H}}$ from 0 to $MN - 1$.

For row $jM + i$ the coefficient of $u_{i,j}$ is given by

$$\begin{aligned}(\mathbf{A}_{\mathbf{H}})_{jM+i,jM+i} = \\ &- \frac{h_y}{h_x}\left[H^{11}(\boldsymbol{s}_{i+1/2,j}) + H^{11}(\boldsymbol{s}_{i-1/2,j})\right] \\ &- \frac{h_x}{h_y}\left[H^{22}(\boldsymbol{s}_{i,j+1/2}) + H^{22}(\boldsymbol{s}_{i,j-1/2})\right].\end{aligned}$$

The four closest neighbours have coefficients

$$
\begin{aligned}
(\mathbf{A_H})_{jM+i,jM+i_p} &= \frac{h_y}{h_x} H^{11}(s_{i-1/2,j}) - \frac{1}{4}\left[H^{12}(\boldsymbol{s}_{i,j+1/2}) - H^{12}(\boldsymbol{s}_{i,j-1/2})\right], \\
(\mathbf{A_H})_{jM+i,jM+i_n} &= \frac{h_y}{h_x} H^{11}(\boldsymbol{s}_{i+1/2,j}) + \frac{1}{4}\left[H^{12}(\boldsymbol{s}_{i,j+1/2}) - H^{12}(\boldsymbol{s}_{i,j-1/2})\right], \\
(\mathbf{A_H})_{jM+i,j_nM+i} &= \frac{h_x}{h_y} H^{22}(\boldsymbol{s}_{i,j+1/2}) + \frac{1}{4}\left[H^{21}(\boldsymbol{s}_{i+1/2,j}) - H^{21}(\boldsymbol{s}_{i-1/2,j})\right], \\
(\mathbf{A_H})_{jM+i,j_pM+i} &= \frac{h_x}{h_y} H^{22}(\boldsymbol{s}_{i,j-1/2}) - \frac{1}{4}\left[H^{21}(\boldsymbol{s}_{i+1/2,j}) - H^{21}(\boldsymbol{s}_{i-1/2,j})\right].
\end{aligned}
$$

Lastly, the four diagonally closest neighbours have coefficients

$$
\begin{aligned}
(\mathbf{A_H})_{jM+i,j_pM+i_p} &= +\frac{1}{4}\left[H^{12}(\boldsymbol{s}_{i,j-1/2}) + H^{21}(\boldsymbol{s}_{i-1/2,j})\right], \\
(\mathbf{A_H})_{jM+i,j_pM+i_n} &= -\frac{1}{4}\left[H^{12}(\boldsymbol{s}_{i,j-1/2}) + H^{21}(\boldsymbol{s}_{i+1/2,j})\right], \\
(\mathbf{A_H})_{jM+i,j_nM+i_p} &= -\frac{1}{4}\left[H^{12}(\boldsymbol{s}_{i,j+1/2}) + H^{21}(\boldsymbol{s}_{i-1/2,j})\right], \\
(\mathbf{A_H})_{jM+i,j_nM+i_n} &= +\frac{1}{4}\left[H^{12}(\boldsymbol{s}_{i,j+1/2}) + H^{21}(\boldsymbol{s}_{i+1/2,j})\right].
\end{aligned}
$$

The rest of the elements of row $jM + i$ are 0.

Based on (S2.10) one can write $\boldsymbol{z} = \mathbf{D}_V^{-1/2}\mathbf{A}\boldsymbol{u}$, where $\mathbf{A} = \mathbf{D}_V\mathbf{D}_{\kappa^2} - \mathbf{A_H}$. This gives the joint distribution of $\boldsymbol{u}$,

$$
\begin{aligned}
\pi(\boldsymbol{u}) &\propto \pi(\boldsymbol{z}) \propto \exp\left(-\frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{z}\right) \\
\pi(\boldsymbol{u}) &\propto \exp\left(-\frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{D}_V^{-1}\mathbf{A}\boldsymbol{u}\right) \\
\pi(\boldsymbol{u}) &\propto \exp\left(-\frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\mathbf{Q}\boldsymbol{u}\right),
\end{aligned}
$$

where $\mathbf{Q} = \mathbf{A}^{\mathrm{T}}\mathbf{D}_V^{-1}\mathbf{A}$. This is a sparse matrix with a maximum of 25 non-zero elements on each row, corresponding to the point itself, its 8 closest neighbours and the 8 closest neighbours of each of the 8 closest neighbours.

S3 Marginal variances with constant coefficients

Proposition S3.1. *Let u be a stationary solution of the SPDE*

$$\kappa^2 u(x,y) - \nabla \cdot \mathbf{H} \nabla u(x,y) = \mathcal{W}(x,y), \qquad (x,y) \in \mathbb{R}^2, \tag{S3.1}$$

where $\mathcal{W}$ is a standard Gaussian white noise process, $\kappa^2 > 0$ is a constant, $\mathbf{H}$ is a positive definite 2×2 matrix and $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$. Then u has marginal variance

$$\sigma_m^2 = \frac{1}{4\pi\kappa^2\sqrt{\det(\mathrm{H})}}.$$

Proof. Since the solution is stationary, Gaussian white noise is stationary, and the SPDE has constant coefficients, the SPDE is acting as a linear filter. Thus one can use spectral theory to find the marginal variance. The transfer function of the SPDE is

$$g(\boldsymbol{w}) = \frac{1}{\kappa^2 + \boldsymbol{w}^{\mathrm{T}}\mathbf{H}\boldsymbol{w}}.$$

The spectral density of a standard Gaussian white noise process on $\mathbb{R}^2$ is identically equal to $1/(2\pi)^2$ so it follows that the spectral density of the solution is

$$f_S(\boldsymbol{w}) = \left(\frac{1}{2\pi}\right)^2 \frac{1}{(\kappa^2 + \boldsymbol{w}^{\mathrm{T}}\mathbf{H}\boldsymbol{w})^2}.$$

From the spectral density it is only a matter of integrating the density over $\mathbb{R}^2$,

$$\sigma_m^2 = \int_{\mathbb{R}^2} f_S(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}.$$

The matrix $\mathbf{H}$ is (symmetric) positive definite and, therefore, has a (symmetric) positive definite square root, say $\mathbf{H}^{1/2}$. Use the change of variables $\boldsymbol{w} = \kappa\mathbf{H}^{-1/2}\boldsymbol{z}$ to find

$$\begin{aligned}
\sigma_m^2 &= \frac{1}{4\pi^2}\int_{\mathbb{R}^2} \frac{1}{(\kappa^2 + \kappa^2\boldsymbol{z}^{\mathrm{T}}\boldsymbol{z})^2}\det(\kappa\mathbf{H}^{-1/2})\,\mathrm{d}\boldsymbol{z} \\
&= \frac{1}{4\pi^2\kappa^2\sqrt{\det(\mathbf{H})}}\int_{\mathbb{R}^2} \frac{1}{(1 + \boldsymbol{z}^{\mathrm{T}}\boldsymbol{z})^2}\,\mathrm{d}\boldsymbol{z} \\
&= \frac{1}{4\pi\kappa^2\sqrt{\det(\mathbf{H})}}.
\end{aligned}$$

□

References

Adler, R.J. and Taylor, J.E. (2007). *Random Fields and Geometry*. Springer Verlag.

Eymard, R., Gallouët, T., and Herbin, R. (2000). Finite Volume Methods. In Ciarlet, P.G. and Lions, J.L., editors, *Solution of Equations in* $\mathbb{R}^n$*, Techniques of Scientific Computing*, 713–1018. Elsevier.

Paper II

Does non-stationary spatial data always require non-stationary random fields?

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H.

2015. *In revision*

Does non-stationary data always require non-stationary random fields?

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

[1]Department of Mathematical Sciences, NTNU, Norway
[2]Department of Mathematical Sciences, University of Bath, United Kingdom

Abstract

A stationary spatial model is an idealization and we expect that the true dependence structures of physical phenomena are spatially varying, but how should we handle this non-stationarity in practice? We study the challenges involved in applying a flexible non-stationary model to a dataset of annual precipitation in the conterminous US, where exploratory data analysis shows strong evidence of a non-stationary covariance structure.

The aim of this paper is to investigate the modelling pipeline once non-stationarity has been detected in spatial data. We show that there is a real danger of over-fitting the model and that careful modelling is necessary in order to properly account for varying second-order structure. In fact, the example shows that sometimes non-stationary Gaussian random fields are not necessary to model non-stationary spatial data.

Keywords: Annual precipitation, Penalized maximum likelihood, Non-stationary Spatial modelling, Stochastic partial differential equations, Gaussian random fields, Gaussian Markov random fields,

1 Introduction

There are, in principle, two sources of non-stationarity present in any dataset: the non-stationarity in the mean and the non-stationarity in the

covariance structure. Classical geostatistical models based on stationary Gaussian random fields (GRFs) ignore the latter, but include the former through covariates that capture important structure in the mean. The focus of non-stationary spatial modelling is non-stationarity in the covariance structure. However, it is impossible to separate the non-stationarity in the mean and the non-stationarity in the covariance structure based on a single realization, and even with multiple realizations it is challenging.

The Karhunen-Loève expansion states that under certain conditions a GRF can be decomposed into an infinite linear combination of orthogonal functions, which is weighted by independent Gaussian variables with decreasing variances. For a single realization these orthogonal functions will be confounded with the covariates in the mean, and the mean structure and the covariance structure cannot be separated. This can give apparent long range dependencies and global non-stationarity if spatial covariates are missing in the mean. Such spurious global non-stationarity and its impact on the local estimation of non-stationarity is an important topic in the paper.

However, the most important point from an applied viewpoint is the computational costs of running a more complex model versus the scientific gain. Non-stationarity in the mean is computationally cheap, whereas methods for non-stationarity in the covariance structure are much more expensive. This raises an important question: How much do we gain by including non-stationarity in the covariance structure? Do we need non-stationary spatial models?

The computational cost of non-stationary models usually comes from a high number of highly dependent parameters that makes it expensive to run MCMC methods or likelihood optimizations, but the challenges with non-stationary models are not only computational. Directly specifying non-stationary covariance functions is difficult and we need other ways of constructing models. Additionally, we need to choose where to put the non-stationarity. Should we have non-stationarity in the range, the anisotropy, the marginal variance, the smoothness or the nugget effect? And how do we combine it all to a valid covariance structure?

1.1 Non-stationarity

Most of the early literature on non-stationary methods deals with data from environmental monitoring stations where multiple realizations are available. In this situation it is possible to calculate the empirical covariances between observed locations, possibly accounting for temporal dependence, and finding the required covariances through, for example, shrinkage towards a parametric model (Loader and Switzer, 1989) or kernel smoothing (Oehlert, 1993). It is also possible to deal efficiently with a single realization with the moving window approach of Haas (Haas, 1990a,b, 1995), but this method does not give valid global covariance structures.

However, the most well-known method from this time period is the deformation method of Sampson and Guttorp (1992), in which an underlying stationary process is made non-stationary by applying a spatial deformation. The original formulation has been extended to the Bayesian framework (Damian et al., 2001, 2003; Schmidt and O'Hagan, 2003), to a single realization (Anderes and Stein, 2008), to covariates in the covariance structure (Schmidt et al., 2011) and to higher dimensional base spaces (Bornn et al., 2012).

Another major class of non-stationary methods is based on the process convolution method developed by Higdon (Higdon, 1998; Higdon et al., 1999). In this method a spatially varying kernel is convolved with a white noise process to create a non-stationary covariance structure. Paciorek and Schervish (2006) looked at a specific case where it is possible to find a closed form expression for a Matérn-like covariance function and Neto et al. (2014) used a kernel that depends on wind direction and strength to control the covariance structure. The process convolution methods have also been extended to dynamic multivariate processes (Calder, 2007, 2008) and spatial multivariate processes (Kleiber and Nychka, 2012).

It is possible to take a different approach to non-stationarity, where instead of modelling infinite-dimensional Gaussian processes one uses a linear combination of basis functions and models the covariance matrix of the coefficients of the basis functions (Nychka et al., 2002, 2014). One such approach is the fixed rank kriging method (Cressie and Johannesson, 2008), which uses a linear combination of a small number of basis functions and estimates the covariance matrix for the coefficients of the linear combination. This approach leads to a continuously indexed spatial

process with a non-stationary covariance structure. The predictive processes (Banerjee et al., 2008) corresponds to a specific choice of the basis functions and the covariance matrix, but does not give a very flexible type of non-stationarity. All such methods are variations of the same concept, but lead to different computational schemes with different computational properties. The dimension of the finite-dimensional basis is in all cases used to control the computational cost and the novelty of each method lies in how the basis elements are selected and connected to each other, and the computational methods used to exploit the structure.

An overview of the literature before around 2010 is given in Sampson (2010). This overview also includes less known methods such as the piecewise Gaussian process of Kim et al. (2005), processes based on weighted linear combination of stationary processes (Fuentes, 2001, 2002a,b; Nott and Dunsmuir, 2002).

Recently, a new class of methods based on the SPDE approach introduced by Lindgren et al. (2011) is emerging. This class of methods is based on a representation of the spatial field as a solution of a stochastic partial differential equation (SPDE) with spatially varying coefficients. The methodology is closely connected with Gaussian Markov random fields (GMRFs) (Rue and Held, 2005) and is able to handle more observations than is possible with the deformation method and the process convolution method. In a similar way as a spatial GMRF describes local behaviour for a discretely indexed process, an SPDE describes local behaviour for a continuously indexed process. This locality in the continuous description can be transferred to a GMRF approximation of the solution of the SPDE, and gives a GMRF with a spatial Markovian structure that can be exploited in computations.

This type of methodology has been applied to global ozone data (Bolin and Lindgren, 2011) and to annual precipitation in Norway with covariates in the covariance structure (Ingebrigtsen et al., 2014). Additionally, Sigrist et al. (2012) used similar type of modelling to handle a spatio-temporal process where wind direction and strength enters in the covariance structure.

Despite all the work that has been done in non-stationary spatial modelling, it is still an open field where no model stands out as the clear choice. However, we believe that modelling locally such as in the SPDE-

based models is more attractive than modelling globally such as in the deformation method and the process convolution method. Therefore, we choose to use an extension of the model by Fuglstad et al. (2015) that allows for both a spatially varying correlation structure and a spatially varying marginal variance. This method is closely connected to the already well-known deformation method of Sampson and Guttorp (1992) and the Matérn-like process convolution of Paciorek and Schervish (2006), but is focused at the local behaviour and not the global behaviour.

In a similar way as in the model of Paciorek and Schervish (2006) the global structure is defined through the combination of ellipses at each location that describe anisotropy. However, their model only combines the ellipses at two and two locations and does not account for the local behaviour between locations. The new model incorporates the local anisotropy everywhere into the covariance for each pair of locations and is not the same as the model of Paciorek and Schervish (2006). The model works in a similar way as the deformation method. However, instead of describing a global deformation, the ellipses augment the local distances around each point and describe locally a change of distances such that lengths are different in different directions, but does not, in general, lead to a deformation of $\mathbb{R}^2$ to $\mathbb{R}^2$. Such local modelling tends to lead to a deformation in an ambient space of dimension higher than 2. The interest of this paper is to study the challenges and results of applying the method to a dataset of annual precipitation in the conterminous US.

1.2 Annual precipitation in the conterminous US

This case study of non-stationarity will use the measurements of monthly total precipitation at different measurement stations in the conterminous US for the years 1895–1997 that are available at `http://www.image.ucar.edu/GSP/Data/US.monthly.met/`. This dataset was chosen because it is publicly available in a form that is easily downloaded and loaded into software, and because the large spatial scale of the dataset and the complexity of the physical process that generates weather makes it intuitively feels like there must be non-stationarity in the dataset.

In total there are 11918 measurement stations in the dataset, but measurements are only available at a subset of the stations each month and the rest of the stations have in-filled data (Johns et al., 2003). For each year,

we aggregate the monthly data at those stations which have measurements available at all months in that year and produce a dataset of yearly total precipitation. This gives a different number of locations for each year. We then take the logarithm of each observation to create the transformed data that is used in this paper. Figure 1 shows the transformed data at the 7040 stations available for 1981. The only covariate available in the dataset is the elevation at each station, and since the focus of the paper is on the covariance structure, no work was done to find other covariates from alternate sources. However, if the focus was to model this data in the best possible way, it would, in general, be good to look for more covariates or consider alternatives such as spatially heterogeneous coefficients before using a full non-stationary model.

We will assume that the transformed data can be treated as Gaussian, which is a reasonable assumption because we are modelling annual precipitation data. However, it would not be a reasonable assumption, for example, for daily data, and it would be necessary to consider not only how to deal with non-stationarity, but also how to deal with the lack of Gaussianity. Bolin and Wallin (2013) compare the predictions made by a stationary Gaussian model, a stationary Gaussian model for transformed data and two stationary non-Gaussian models for monthly precipitation for two different months from the same dataset as in this paper. They apply the non-stationary model of Bolin (2014), but do not find clear evidence that one model perform better than the others. The approach of Bolin (2014) is built on the same principles as the approach in this paper and a possible extension of the presented non-stationary model would be to non-Gaussian data.

The main motivation for focusing on the year 1981 is that Paciorek and Schervish (2006) previously studied the annual precipitation in the subregion of Colorado for this year. They did not see major improvements over a stationary model and our preliminary analysis showed that there was little non-stationarity left in the subregion after introducing a joint mean and elevation. However, Colorado constitutes a small part of the conterminous US, and as shown in Figure 2 there are large differences in the topography of the western and the eastern part of the conterminous US. A large proportion of the western part is mountainous whereas in the eastern part a large proportion is mostly flat. This varied topography is a

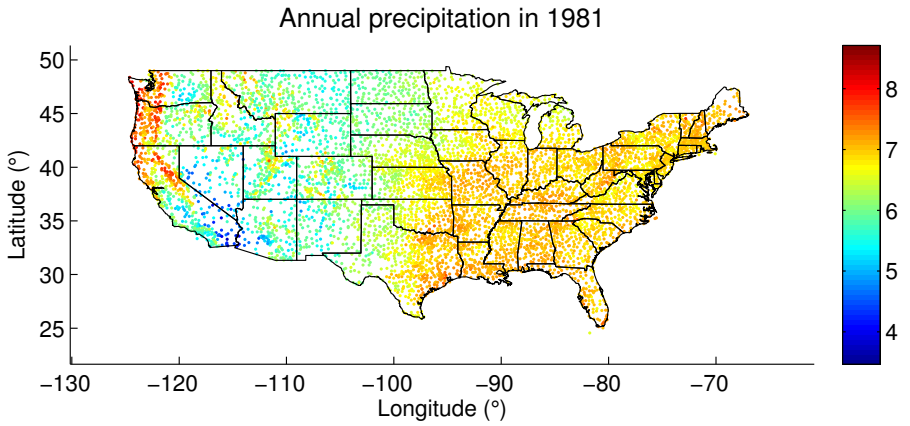


Figure 1: The logarithm of total yearly precipitation measured in millimetres at 7040 locations in the conterminous US for the year 1981.

strong indication that the process cannot possibly be stationary.

To substantiate our claims of non-stationarity we explore the difference in the covariance structure in the western and eastern part through variograms. The data from years 1971–1985 is selected and divided into two regions: longitude less than 100 °W and longitude greater than or equal to 100 °W. For each year the variogram of each region is calculated. Figure 3 shows that there is no overlap between the variograms of the western region and the eastern region. There is significant variation within each region, but the overall appearance clearly indicates different covariance structures within the regions. Based on the evidence of non-stationarity seen in the variograms for the full region, we want to know if a non-stationary model will improve the predictions. It has been observed by several authors (Schmidt et al., 2011; Neto et al., 2014) and it has also been the experience of the authors that non-stationary models do not lead to much difference in the predicted values, and that the differences are found in the prediction variances. However, predictions should always have associated error estimates and when we write improved predictions, we are interested in whether the predictive distributions, summarized by the predicted values and their associated prediction variances, better describe the observed values.

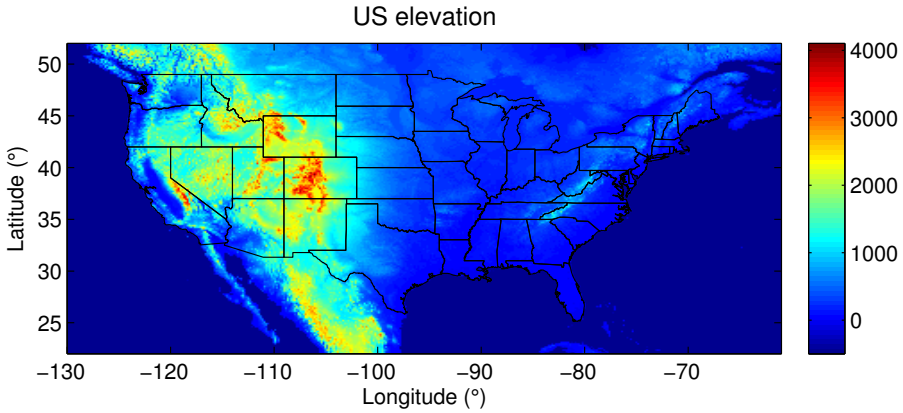


Figure 2: Elevation in the US measured in meters. Data from GLOBE data set (Hastings et al., 1999)

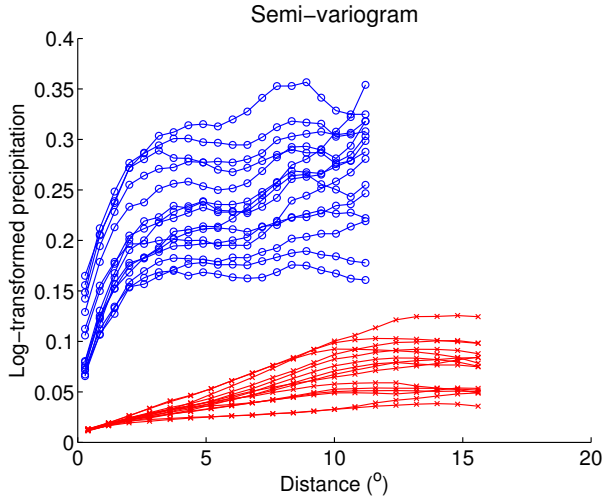


Figure 3: Estimated semi-variograms for the years 1971 to 1985 using the locations with longitudes less than 100 °W coloured in blue and marked with circles and with longitudes greater than 100 °W coloured in red and marked with crosses.

There are two cases of interest: a single realization and multiple realizations. In the former it is impossible to separate the non-stationarity in the mean and in the covariance structure, and the non-stationary model might be more accurately described as adaptive smoothing, but many spatial datasets are of this form and a non-stationary model might still perform better than a stationary model. We will investigate both of these cases and evaluate whether the non-stationary model improves predictions and whether the computational costs are worth it. It is clear that stationarity is not the truth, but that does not mean that it does not necessarily constitutes a sufficient model for predictions.

1.3 Overview

The paper is divided into five sections. Section 2 describes how we model the data. We discuss what type of non-stationarity is present in the model and how it is specified, how we parametrize the non-stationarity and how we perform computations with the non-stationary model. Then in Section 3 a hierarchical model incorporating the non-stationary model is applied to annual precipitation in a single realization setting, and in Section 4 the data is studied from a multiple realizations perspective. The differences between the estimated covariance structures and the prediction scores for the different models are discussed. The paper ends with discussion and concluding remarks in Section 5.

2 Modelling the data

Before analyzing the data we need to introduce the model that will be used. Particularly, we need to say which types of non-stationarity that will be present in the model and how this non-stationarity will be modelled. A good spatial model should provide a useful way to do both the theoretical modelling and the associated computations. We first discuss the theoretical part, and then discuss how to do the computations and how to parametrize the non-stationary.

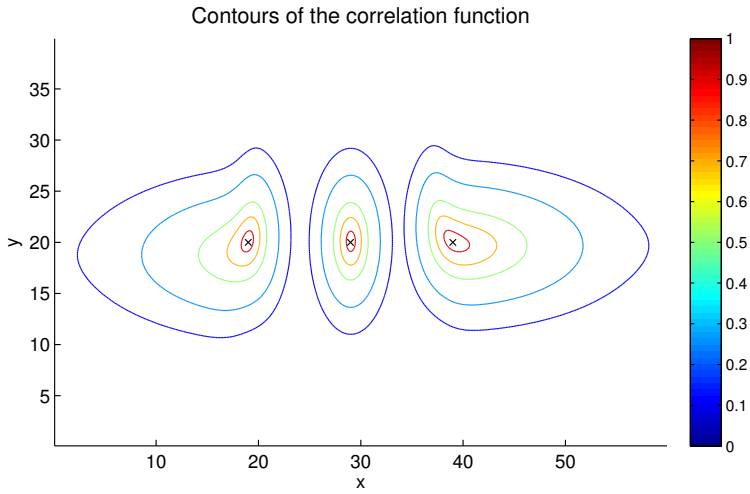


Figure 4: Example of a correlation function caused by varying local behaviour. For each location marked with a black cross, the 0.9, 0.7, 0.5, 0.25 and 0.12 level contours of the correlation function are shown.

2.1 Modelling the non-stationarity

It is difficult to specify a global covariance function when one only has intuition about local behaviour. Consider the situation in Figure 4. The left hand side and the right hand side have locally large "range" in the horizontal direction and somewhat shorter "range" in the vertical direction, and the middle area has locally much shorter "range" in the horizontal direction, but slightly longer in the vertical direction. We write "range" with quotation marks because the concept of a global range does not have a well-defined meaning in non-stationary modelling. Instead we will think of range as a local feature and use the word to mean what happens to dependency in a small region around each point. From the figure one can see that for the point in the middle, the chosen contours look more or less unaffected by the two other regions since they are fully contained in the middle region, but that for the point on the left hand side and the point on the right hand side, there is much skewness introduced by the transition into a different region.

It would be hard to specify a fitting global correlation function for

this situation. However, if one instead starts with an isotropic process and then stretches the left hand side and the right hand side in the x-direction, the task is much easier. This is a flexible way to create interesting covariance structures and is the core of the deformation method (Sampson and Guttorp, 1992), but can be challenging since one has to create a valid *global* deformation. We present instead a model where the modelling can be done *locally* without worrying about the *global* structure. We let the local structure automatically specify a valid global structure. In this example one would only specify that locally the range is longer in the horizontal direction in the left hand side and the right hand side, and then let this implicitly define the global structure without directly modelling a global deformation.

In the SPDE-based approach the correlation between two spatial locations is determined implicitly by the behaviour between the spatial locations. If there are mountains, the model could specify that locally the distances are longer than they appear on the map and the correlation will decrease more quickly when crossing those areas, and if there are plains, the model could specify that distances are shorter than they appear on the map and the correlation will decrease more slowly in those areas. A major advantage of this approach is that the local specification naturally leads to a spatial GMRF with good computational properties. It is possible to approximate the local continuous description with a local discrete description. The result is a spatial GMRF with a very sparse precision matrix

The starting point for the non-stationary SPDE-based model is the stationary SPDE introduced in Lindgren et al. (2011),

$$(\kappa^2 - \nabla \cdot \nabla)u(\vec{s}) = \sigma \mathcal{W}(\vec{s}), \qquad \vec{s} \in \mathbb{R}^2, \tag{1}$$

where $\kappa > 0$ and $\sigma > 0$ are constants, $\nabla = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y})^{\mathrm{T}}$ and $\mathcal{W}$ is a standard Gaussian white noise process. The SPDE describes the GRF u as a smoothed version of the Gaussian white noise on the right hand side of the equation. Whittle (1954, 1963) showed that any stationary solution of this SPDE has the Matérn covariance function

$$r(\vec{s}_1, \vec{s}_2) = \frac{\sigma^2}{4\pi\kappa^2}(\kappa||\vec{s}_2 - \vec{s}_1||)K_1(\kappa||\vec{s}_2 - \vec{s}_1||), \tag{2}$$

where K_1 is the modified Bessel function of second kind, order 1. This covariance function is a member of the commonly-used Matérn family of covariance functions, and one can see from Equation (2) that one can first use κ to select the range and then σ to achieve the desired marginal variance. In some methods for non-stationarity it is possible to spatially vary the smoothness, but this is not a feature that is available in the non-stationary model presented here. However, with the flexibility present in the rest of the non-stationarity it is not clear if the smoothness would be jointly identifiable.

The next step is to generate a GRF with an anisotropic Matérn covariance function. The cause of the isotropy in SPDE (1) is that the Laplacian, $\Delta = \nabla \cdot \nabla$ is invariant to a change of coordinates that involves rotation and translation. To change this a 2×2 matrix $\mathbf{H} > 0$ is introduced into the operator to give the SPDE

$$(\kappa^2 - \nabla \cdot \mathbf{H} \nabla) u(\vec{s}) = \sigma \mathcal{W}(\vec{s}). \tag{3}$$

This choice is closely related to the change of coordinates $\tilde{\vec{s}} = \mathbf{H}^{1/2}\vec{s}$ (see Fuglstad et al. (2015, Section 3)) and gives the covariance function

$$r(\vec{s}_1, \vec{s}_2) = \frac{\sigma^2}{4\pi\kappa^2\sqrt{\det(\mathbf{H})}}(\kappa||\mathbf{H}^{-1/2}(\vec{s}_2 - \vec{s}_1)||)K_1(\kappa||\mathbf{H}^{-1/2}(\vec{s}_2 - \vec{s}_1)||). \tag{4}$$

Compared to Equation (2) there is a change in the marginal variance and a directionality is introduced through a distance measure different than the standard Euclidean distance. Figure 5 shows how the eigenpairs of $\mathbf{H}$ and the value of κ act together to control range. One can see that the construction leads to elliptic iso-covariance curves. In what follows σ is assumed to be equal to 1 since the marginal variance can be controlled by varying κ^2 and $\mathbf{H}$ together.

The final step is to construct a non-stationary GRF where the local behaviour at each location is governed by SPDE (3) with $\sigma = 1$ and the values of κ^2 and $\mathbf{H}$ varying over the domain. The intention is to create a GRF by chaining together processes with different local covariance structures. The SPDE becomes

$$(\kappa^2(\vec{s}) - \nabla \cdot \mathbf{H}(\vec{s}) \nabla) u(\vec{s}) = \mathcal{W}(\vec{s}). \tag{5}$$

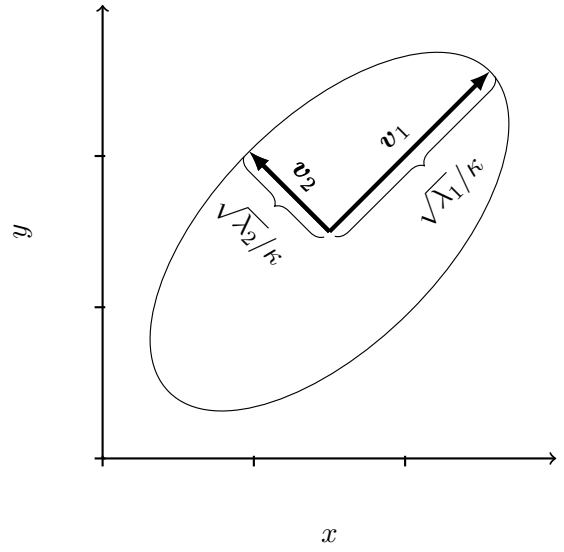


Figure 5: Iso-correlation curve for the 0.6 level, where $(\lambda_1, \vec{v}_1)$ and $(\lambda_2, \vec{v}_2)$ are the eigenpairs of $\mathbf{H}$.

For technical reasons concerned with the discretization in the next section, κ^2 is required to be continuous and $\mathbf{H}$ is required to be continuously differentiable. This does not present any problems and is easily achieved by using continuously differentiable basis functions for κ^2 and $\mathbf{H}$. The restricted form where κ^2 is constant was investigated in Fuglstad et al. (2015), but this restricted form only allows for varying local anisotropy without control over the marginal variances. This extended model allows for spatially varying "range", anisotropy and marginal variance.

2.2 Discrete model for computations

SPDE (5) describes the covariance structure of a GRF, but before the model can be used in practice the description must be brought into a form which is useful for computations. The first thing to notice is that the operator in front of u only contains multiplications with functions and first order and second order derivatives. All of these operations involve only the local properties of u at each location. This means that if u is dis-

cretized using a finite-dimensional local basis expansion, the corresponding discretized operators (matrices) should only involve variables close to each other. This can be exploited to create a sparse GMRF which possesses approximately the same covariance structure as u. The arguments above are not applicable for all smoothnesses, but we are constructing a model where the smoothness is fixed to 1 and the range is allowed to vary spatially (See discussion in Fuglstad et al. (2015, p. 5)). A detailed description of the basis function expansion, the choice of mesh, and the theoretical properties of the methods described in this section in Lindgren et al. (2011); Simpson et al. (2012, 2011).

The first step in creating the GMRF is to restrict SPDE (5) to a bounded domain,

$$(\kappa^2(\vec{s}) - \nabla \cdot \mathbf{H}(\vec{s})\nabla)u(\vec{s}) = \mathcal{W}(\vec{s}), \qquad \vec{s} \in \mathcal{D} = [A_1, B_1] \times [A_2, B_2] \subset \mathbb{R}^2,$$

where $B_1 > A_1$ and $B_2 > A_2$. This restriction necessitates a boundary condition to make the distribution useful and proper. For technical reasons the boundary condition chosen is zero flux across the boundaries, i.e. at each point of the boundary the flux $\mathbf{H}(\vec{s})\nabla_{\vec{n}}u(\vec{s})$, where $\vec{n}$ is the normal vector of the boundary at that point, is zero. The derivation of a discretized version of this SPDE on a grid is involved, but for periodic boundary conditions the derivation can be found in the supplementary material to Fuglstad et al. (2015). The boundary conditions in this problem involve only a slight change in that derivation.

For a regular $m \times n$ grid of $\mathcal{D}$, the end result is the matrix equation

$$\mathbf{A}(\kappa^2, \mathbf{H})\vec{u} = \frac{1}{\sqrt{V}}\vec{z},$$

where V is the area of each cell in the grid, $\vec{u}$ corresponds to the values of u on the cells in the regular grid stacked column-wise, $\vec{z} \sim \mathcal{N}_{mn}(\vec{0}, \mathbf{I}_{mn})$ and $\mathbf{A}(\kappa^2, \mathbf{H})$ is a discretized version of $(\kappa^2 - \nabla \cdot \mathbf{H}\nabla)$. This matrix equation leads to the multivariate Gaussian distribution

$$\vec{u} \sim \mathcal{N}_{mn}(\vec{0}, \mathbf{Q}(\kappa^2, \mathbf{H})^{-1}), \tag{6}$$

where $\mathbf{Q}(\kappa^2, \mathbf{H}) = \mathbf{A}(\kappa^2, \mathbf{H})^{\mathrm{T}}\mathbf{A}(\kappa^2, \mathbf{H})V$. The precision matrix $\mathbf{Q}$ is proper and has up to 25 non-zero elements in each row, corresponding to the point

itself, its eight closest neighbours and the eight closest neighbours of each of the eight closest neighbours. Since the approximation is constructed from an SPDE, it behaves consistently over different resolution and converges to a continuously indexed model for small resolutions. Changing the resolution changes which features can be represented by the model, but does not induce large changes to the covariance structure.

This construction alleviates one of the largest problems with GMRFs, namely that they are hard to specify in a spatially coherent manner. The computational benefits of spatial GMRFs are well known, but a GMRF needs to be constructed through its conditional distributions and it is notoriously hard to do this for non-stationary models. But with the derivation outlined above it is possible to model the problem with an SPDE and then do the computations with the computational benefits of a spatial GMRF.

2.3 Parametrizing the non-stationarity

Before we can turn the theoretical and computational description of the non-stationary model into a statistical model, we need to describe the non-stationarity through parameters. This means both decomposing the model into parameters and connecting the parameters together through a penalty.

The first step is to decompose the function $\mathbf{H}$, which must give positive definite 2×2 matrices at each location, into simpler functions. One usual way to do this is to use two strictly positive functions λ_1 and λ_2 for the eigenvalues and a function ϕ for the angle between the x-axis and the eigenvector associated with λ_1. However, with a slight re-parametrization $\mathbf{H}$ can be written as the sum of an isotropic effect, described by a constant times the identity matrix, plus an additional anisotropic effect, described by direction and magnitude.

Express $\mathbf{H}$ through the scalar functions γ, v_x and v_y by

$$\mathbf{H}(\vec{s}) = \gamma(\vec{s})\mathbf{I}_2 + \begin{bmatrix} v_x(\vec{s}) \\ v_y(\vec{s}) \end{bmatrix} \begin{bmatrix} v_x(\vec{s}) & v_y(\vec{s}) \end{bmatrix},$$

where γ is required to be strictly positive. The eigendecomposition of this matrix has eigenvalue $\lambda_1(\vec{s}) = \gamma(\vec{s}) + v_x(\vec{s})^2 + v_y(\vec{s})^2$ with eigenvector $\vec{v}_1(\vec{s}) = (v_x(\vec{s}), v_y(\vec{s}))$ and eigenvalue $\lambda_2(\vec{s}) = \gamma(\vec{s})$ with eigenvector $\vec{v}_2(\vec{s}) = (-v_y(\vec{s}), v_x(\vec{s}))$. From Figure 5 this means that for a stationary model, γ

affects the length of the shortest semi-axis of the iso-correlation curves and $\vec{v}$ specifies the direction of and how much larger the longest semi-axis is. The above decomposition through γ, v_x and v_y is general and is valid for every symmetric positive-definite 2×2 matrix.

Since we want flexible covariance structures, some representation of the functions κ^2, γ, v_x and v_y is needed. To ensure positivity of κ^2 and γ, they are first transformed into $\log(\kappa^2)$ and $\log(\gamma)$. Each of these functions will be expanded in a basis, and requires a penalty that imposes regularity and makes sure the function is not allowed to vary too much. The choice was made to give $\log(\kappa^2)$, $\log(\gamma)$, v_x and v_y spline-like penalties. The steps that follow are the same for each function. Therefore, they are only shown for $\log(\kappa^2)$.

The function $\log(\kappa^2)$ is given a penalty according to the distribution generated from the SPDE

$$-\Delta \log(\kappa^2(\vec{s})) = \mathcal{W}_\kappa(\vec{s})/\sqrt{\tau_\kappa}, \qquad \vec{s} \in \mathcal{D}, \tag{7}$$

where $\tau_\kappa > 0$ is the parameter controlling the penalty, with the Neumann boundary condition of zero derivatives at the edges. This extra requirement is used to restrict the resulting distribution so it is only invariant to the addition of a constant function, and the penalty parameter is used to control how much $\log(\kappa^2)$ can vary from a constant function. The penalty defined through SPDE (7) is in this paper called a two-dimensional second-order random walk due to its similarity to a one-dimensional second-order random walk (Lindgren and Rue, 2008).

The first step of making the above penalty applicable for the computational model is to expand $\log(\kappa^2)$ in a basis through a linear combination of basis functions,

$$\log(\kappa^2(\vec{s})) = \sum_{i=1}^{k} \sum_{j=1}^{l} \alpha_{ij} f_{ij}(\vec{s}),$$

where $\{\alpha_{ij}\}$ are the parameters and $\{f_{ij}\}$ are real-valued basis functions. For convenience, the basis is chosen in such a way that all basis functions satisfy the boundary conditions specified in SPDE (7). If this is done, one immediately satisfies the boundary condition. The remaining tasks are then to decide which basis functions to use and what the resulting penalties on the parameters are.

Due to a desire to make $\mathbf{H}$ continuously differentiable and a desire to have "local" basis functions, the basis functions are chosen to be based on 2-dimensional, second-order B-splines (piecewise-quadratic functions). The basis is constructed as a tensor product of two 1-dimensional B-spline bases constrained to satisfy the boundary condition.

The penalty is based on the distribution defined by SPDE (7), so the final step is to determine a Gaussian distribution for the parameters such that the distribution of $\log(\kappa^2)$ is close to a solution of SPDE (7). The approach taken is based on a least-squares formulation of the solution and is described in Appendix A. Let $\vec{\alpha}$ be the $\{\alpha_{ij}\}$ parameters stacked row-wise, then the result is that α should be given a zero-mean Gaussian distribution with precision matrix $\tau_\kappa \mathbf{Q}_{\mathrm{RW2}}$. This matrix has rank $(kl-1)$, due to the Neumann boundary conditions, and the distribution is invariant to the addition of a vector of only the same values, but for convenience the penalty will still be written as $\vec{\alpha} \sim \mathcal{N}_{kl}(\vec{0}, \mathbf{Q}_{\mathrm{RW2}}^{-1}/\tau_\kappa)$.

2.4 Hierarchical model

Observations $y_1, y_2, \ldots, y_N$ are made at locations $\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_N$. The observed value at each location is assumed to be the sum of a fixed effect due to covariates, a spatial "smooth" effect and a random effect. The covariates at location $\vec{s}_i$ are described by the p-dimensional row vector $\vec{x}(\vec{s}_i)^{\mathrm{T}}$ and the spatial field is denoted by u. This gives the observation equation

$$y_i = \vec{x}(\vec{s}_i)^{\mathrm{T}}\vec{\beta} + u(\vec{s}_i) + \epsilon_i,$$

where $\vec{\beta}$ is a p-variate random vector for the coefficients of the covariates and $\epsilon_i \sim \mathcal{N}(0, 1/\tau_{\mathrm{noise}})$ is the random effect for observation i, for $i = 1, 2, \ldots, N$.

The u is modelled and parametrized as described in the previous sections and the GMRF approximation is used for computations. In this GMRF approximation the domain is divided into a regular grid consisting of rectangular cells and each element of the GMRF approximation describes the average value on one of these cells. So $u(\vec{s}_i)$ is replaced with the approximation $\vec{e}(\vec{s}_i)^{\mathrm{T}}\vec{u}$, where $\vec{e}(\vec{s}_i)^{\mathrm{T}}$ is the mn-dimensional row vector selecting the element of $\vec{u}$ which corresponds to the cell which contains location $\vec{s}_i$. In total, this gives

$$\vec{y} = \mathbf{X}\vec{\beta} + \mathbf{E}\vec{u} + \vec{\epsilon}, \tag{8}$$

where $\vec{y} = (y_1, y_2, \ldots, y_N)$, the matrix $\mathbf{X}$ has $\vec{x}(\vec{s}_1)^{\mathrm{T}}, \ldots, \vec{x}(\vec{s}_N)^{\mathrm{T}}$ as rows and the matrix $\mathbf{E}$ has $\vec{e}(\vec{s}_1)^{\mathrm{T}}, \ldots, \vec{e}(\vec{s}_N)^{\mathrm{T}}$ as rows. In this equation the spatial effect is approximated with a discrete model, but the covariate has not been gridded and is at a higher resolution than the grid.

The model for the observations can also be written in the form

$$\vec{y}|\vec{\beta}, \vec{u}, \log(\tau_{\text{noise}}) \sim \mathcal{N}_N(\mathbf{X}\vec{\beta} + \mathbf{E}\vec{u}, \mathbf{I}_N/\tau_{\text{noise}}).$$

The parameter τ_{noise} acts as the precision of a joint effect from measurement noise and small scale spatial variation (Diggle et al., 2007). We make the underlying model for the p-dimensional random variable $\vec{\beta}$ proper by introducing a weak Gaussian penalty,

$$\vec{\beta} \sim \mathrm{N}_p(\vec{0}, \mathbf{I}_p/\tau_\beta).$$

The penalty can be made stronger, but we do not believe it will have a strong effect on the estimates for this dataset with only an intercept and one covariate.

To describe the full hierarchical model, we introduce symbols to denote the parameters that control the spatial field u. Denote the parameters that control $\log(\kappa^2)$, $\log(\gamma)$, v_x and v_y by $\vec{\alpha}_1$, $\vec{\alpha}_2$, $\vec{\alpha}_3$ and $\vec{\alpha}_4$, respectively. Further, denote the corresponding penalty parameters for each function by τ_1, τ_2, τ_3 and τ_4. With this notation the full model becomes

$$\begin{aligned}
&\text{Stage 1: } \vec{y}|\vec{\beta}, \vec{u}, \log(\tau_{\text{noise}}) \sim \mathcal{N}_N(\mathbf{X}\vec{\beta} + \mathbf{E}\vec{u}, \mathbf{I}_N/\tau_{\text{noise}}) \\
&\text{Stage 2: } \vec{u}|\vec{\alpha}_1, \vec{\alpha}_2, \vec{\alpha}_3, \vec{\alpha}_4 \sim \mathcal{N}_{nm}(\vec{0}, \mathbf{Q}^{-1}), \quad \vec{\beta} \sim \mathcal{N}_p(\vec{0}, \mathbf{I}_p/\tau_\beta) \\
&\text{Stage 3: } \vec{\alpha}_i|\tau_i \sim \mathcal{N}_{kl}(\vec{0}, \mathbf{Q}_{\text{RW2}}^{-1}/\tau_i) \text{ for } i = 1, 2, 3, 4,
\end{aligned}$$

where τ_1, τ_2, τ_3, τ_4 and τ_β are penalty parameters that must be preselected.

An important model choice when constructing the GMRF approximation of the spatial process is the selection of the resolution of the approximation. The approximation does not allow for variation of the spatial field within a grid cell and the spatial resolution must be chosen high enough to capture variations on the scale at which observations were made. The variation at sub-grid scale cannot be captured by the approximation and will be captured by the nugget effect.

2.5 Penalized likelihood and inference

The two things of main interest to us in this case study are the covariance parameters $\vec{\theta} = (\vec{\alpha}_1, \vec{\alpha}_2, \vec{\alpha}_3, \vec{\alpha}_4, \log(\tau_{\text{noise}}))$ and the predictive distributions for unmeasured locations. To estimate the covariance parameters, we need the integrated likelihood where the latent field consisting of the coefficients of the fixed effects and the spatial effect are integrated out. This integration can be done explicitly because the spatial field by construction is Gaussian and the parameters of the fixed effects are Gaussian due to the choice of a Gaussian penalty.

First, collect the fixed effect and the spatial effect in $\vec{z} = (\vec{u}^{\mathrm{T}}, \vec{\beta}^{\mathrm{T}})$. The model given the value of $\vec{\theta}$ can then be written as

$$\vec{z}|\vec{\theta} \sim \mathcal{N}_{mn+p}(\vec{0}, \mathbf{Q}_z^{-1})$$

and

$$\vec{y}|\vec{z}, \vec{\theta} \sim \mathcal{N}_N(\mathbf{S}\vec{z}, \mathbf{I}_N/\tau_{\text{noise}}),$$

where

$$\mathbf{S} = \begin{bmatrix} \mathbf{E} & \mathbf{X} \end{bmatrix} \quad \text{and} \quad \mathbf{Q}_z = \begin{bmatrix} \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \tau_\beta \mathbf{I}_p \end{bmatrix}.$$

We then use the fact that both these distributions are Gaussian to integrate out $\vec{z}$ from the likelihood, as shown in Appendix B. This gives the full penalized log-likelihood

$$\begin{aligned} &\log(\pi(\vec{\theta}|\vec{y})) \\ &\quad = \text{Const} - \frac{1}{2}\sum_{i=1}^{4} \vec{\alpha}_i^{\mathrm{T}} \mathbf{Q}_{\text{RW2}} \vec{\alpha}_i \cdot \tau_i + \frac{1}{2}\log(\det(\mathbf{Q}_z)) + \frac{N}{2}\log(\tau_{\text{noise}}) + \\ &\quad - \frac{1}{2}\log(\det(\mathbf{Q}_{\text{C}})) - \frac{1}{2}\vec{\mu}_{\text{C}}^{\mathrm{T}} \mathbf{Q}_z \vec{\mu}_{\text{C}} - \frac{\tau_{\text{noise}}}{2}(\vec{y} - \mathbf{S}\vec{\mu}_{\text{C}})^{\mathrm{T}}(\vec{y} - \mathbf{S}\vec{\mu}_{\text{C}}), \end{aligned} \tag{9}$$

where $\mathbf{Q}_{\text{C}} = \mathbf{Q}_z + \mathbf{S}^{\mathrm{T}}\mathbf{S} \cdot \tau_{\text{noise}}$ and $\vec{\mu}_{\text{C}} = \mathbf{Q}_{\text{C}}^{-1}\mathbf{S}^{\mathrm{T}}\vec{y} \cdot \tau_{\text{noise}}$.

The first step of the inference scheme is to estimate the covariance parameters $\vec{\theta}$ with the value $\hat{\vec{\theta}}$ that maximizes Equation (9). This value is then used to calculate predictions and prediction standard deviations at new locations $\vec{y}^*$ by using the predictive distribution $\vec{y}^*|\hat{\vec{\theta}}, \vec{y}$. However, the

penalty parameters that control the penalty of the covariance parameters are difficult to estimate. The profile likelihoods are hard to calculate and there is not enough information on such a low stage of the hierarchical model to estimate them together with the covariance parameters. Thus they have to be pre-selected, based on intuition about how much the covariance structure should be allowed to vary, or chosen with a cross-validation procedure based on a scoring rule for the predictions.

During implementation of the inference scheme it became apparent that an analytic expression for the gradient was needed for the optimization to converge. Its form is given in Appendix C, and its value can be computed for less cost than a finite difference approximation of the gradient for the number of parameters used in the application in this paper. The calculations require the use of techniques for calculating only parts of the inverse of a sparse precision matrix (Rue and Held, 2010).

3 Non-stationarity in a single realization

3.1 Adaptive smoothing framework

We begin by considering the common situation in spatial statistics where only a single realization is available. In this situation it is theoretically impossible to separate non-stationarity in the mean and in the covariance structure, and the non-stationary model is better described as adaptive smoothing. The non-stationary model allows the degree of smoothing to vary over space, and areas with long range will have high smoothing and areas with short range will have low smoothing. The non-stationary model will necessarily include part of the non-stationarity in the mean in the covariance structure, but this is not necessarily a problem and might lead to better predictions. The main interest is finding out whether the complex non-stationary model improves predictions at unobserved locations and whether the computational costs are worth it.

We select the year 1981 which has 7040 measurement stations and want to predict the annual precipitation in the entire conterminous US with associated prediction standard deviations. Two covariates are used: a joint mean and elevation. This means that the design matrix, $\mathbf{X}$, in Equation (8) has two columns. The first column contains only ones, and

Table 1: Estimated values of the parameters and associated approximate standard deviations for the stationary model.

Parameter	**Estimate**	**Standard deviation**
$\log(\kappa^2)$	-1.75	0.15
$\log(\gamma)$	-0.272	0.042
v_x	0.477	0.053
v_y	-0.313	0.057
$\log(\tau_{\text{noise}})$	4.266	0.030

corresponds to the joint mean, and the second column contains elevations measured in kilometres. There should be strong information about the two covariates and a weak penalty is applied to the coefficients of the fixed effects, $\vec{\beta} \sim \mathcal{N}_2(\vec{0}, \mathbf{I}_2 \cdot 10^4)$.

3.2 Stationary model

The spatial effect is constructed on a rectangular domain with longitudes from 130.15 °W to 60.85 °W and latitudes from 21.65 °N to 51.35 °N. This is larger than the actual size of the conterminous US as can be seen in Figure 1, and is chosen to reduce boundary effects. The domain is discretized into a 400×200 grid and the parameters $\log(\kappa^2)$, $\log(\gamma)$, v_x, v_y and $\log(\tau_{\text{noise}})$ are estimated. In this case the second order random walk penalty is not used as no basis (except a constant) is needed for the functions. The estimated values with associated approximate standard deviations are shown in Table 1. The approximate standard deviations are calculated from the observed information matrix.

From Section 2.1 one can see that the estimated model implies a covariance function approximately equal to the Matérn covariance function

$$r(\vec{s}_1, \vec{s}_2) = \hat{\sigma}^2 \left|\left| \left(\hat{\mathbf{H}}/\hat{\kappa}^2\right)^{-1/2} (\vec{s}_2 - \vec{s}_1) \right|\right| K_1 \left(\left|\left| \left(\hat{\mathbf{H}}/\hat{\kappa}^2\right)^{-1/2} (\vec{s}_2 - \vec{s}_1) \right|\right| \right),$$

where $\hat{\sigma}^2 = 0.505$ and

$$\frac{\hat{\mathbf{H}}}{\hat{\kappa}^2} = \begin{bmatrix} 5.71 & -0.86 \\ -0.86 & 4.96 \end{bmatrix},$$

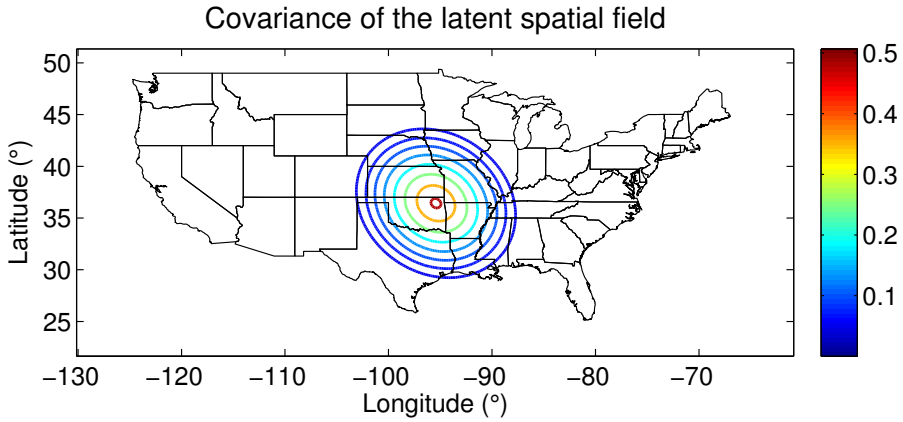


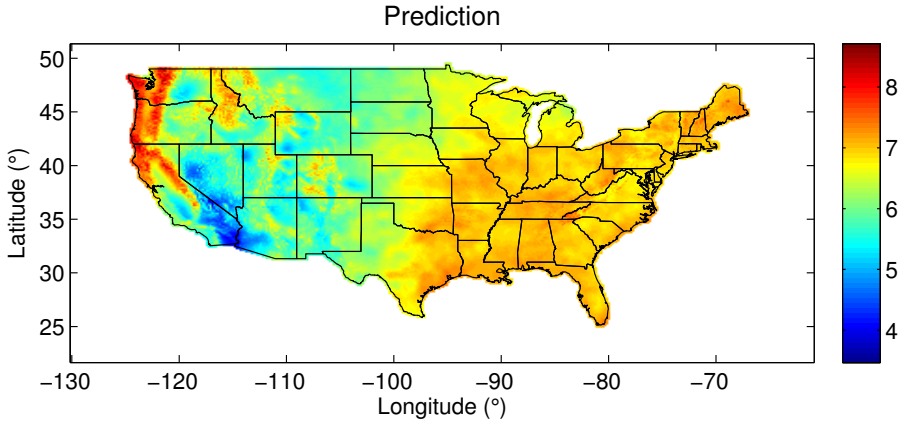
Figure 6: The 0.95, 0.70, 0.50, 0.36, 0.26, 0.19, 0.14 and 0.1 level correlation contours of the estimated covariance function for the stationary model.

together with a nugget effect with precision $\hat{\tau}_{\text{noise}} = 71.2$. Figure 6 shows contours of the estimated covariance function with respect to a chosen location. One can see that the model gives high dependence within a typical-sized state, whereas there is little dependence between the centres of different typically-sized states.

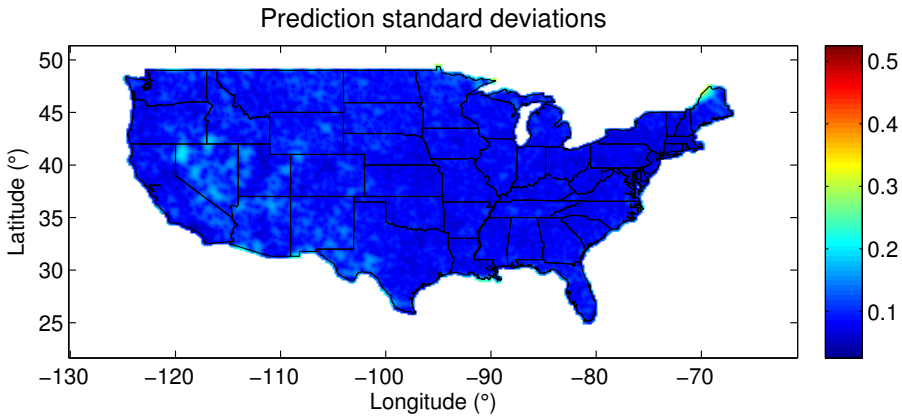
Next, the parameter values are used together with the observed logarithms of annual precipitations to predict the logarithm of annual precipitation at the centre of each cell in the discretization. The elevation covariate for each location is selected from bilinear interpolation from the closest points in the high resolution elevation data set GLOBE (Hastings et al., 1999). The predictions and prediction standard deviations are shown in Figure 7. Since there only are observations within the conterminous US and this is the area of interest, the locations outside are coloured white.

3.3 Non-stationary model

The parameters τ_1, τ_2, τ_3 and τ_4, that appear in the penalty for the functions $\log(\kappa^2)$, $\log(\gamma)$, v_x and v_y, respectively, have to be chosen before the rest of the inference is started. The parameters are chosen with 5-fold cross-validation based on the log-predictive density. The data is randomly



(a) Prediction



(b) Prediction standard deviations

Figure 7: Predicted values and prediction standard deviations for the stationary model.

divided into five parts and in turn one part is used as test data and the other four parts are used as training data. For each choice of τ_1, τ_2, τ_3 and τ_4 the cross-validation error is calculated by

$$\text{CV}(\tau_1, \tau_2, \tau_3, \tau_4) = -\frac{1}{5}\sum_{i=1}^{5} \log(\pi(\vec{y}_i^*|\vec{y}_i, \hat{\vec{\theta}}_i),$$

where $\vec{y}_i^*$ is the test data and $\hat{\vec{\theta}}_i$ is the estimated covariance parameters based on the training data $\vec{y}_i$ using the selected τ-values. The cross validation is done over $\log(\tau_i) \in \{2, 4, 6, 8\}$ for $i = 1, 2, 3, 4$. We selected four values for each parameter to have a balance between the need to test strong and weak penalties and to make the problem computationally feasible. Controlling the penalty on non-stationarity is important, but appropriate penalty values are not easily deduced from the model. Therefore, different values were tested to determine values of τ_i that corresponds to a weak penalty and a strong penalty and then four points were chosen linearly on log-scale since τ_i acts as a scale parameter. We use the same domain size as for the stationary model, but reduce the grid size to 200×100 with 8×4 basis functions for each function. The choice that gave the smallest cross-validation error was $\log(\tau_1) = 2$, $\log(\tau_2) = 4$, $\log(\tau_3) = 2$ and $\log(\tau_4) = 8$.

After the penalty parameters are selected, the grid size is increased to 400×200 and each of the four functions in the SPDE is given a 16×8 basis functions. Together with the precision parameter of the random effect this gives a total of 513 parameters. These parameters are estimated together based on the integrated likelihood. Note that there are not 513 "free" parameters as they are connected together in four different penalties enforcing slowly changing functions. This means that an increase in the number of parameters increases the resolutions of the functions, but not directly the degree of freedom in the model.

The nugget effect is estimated to have a precision of $\hat{\tau}_{\text{noise}} = 107.4$. The estimates of κ^2 and $\mathbf{H}$ are not shown since the exact values themselves are not interesting. We calculate instead the marginal standard deviations for all locations and 0.7 level correlation contours for selected locations in Figure 8(a) and Figure 8(b), respectively. From these figures one can see that the estimated covariance structure is different from the estimated

covariance structure for the stationary model shown in Figure 6. In the non-stationary model we have a much longer range in the eastern part and a much short range in the mountainous areas in the west.

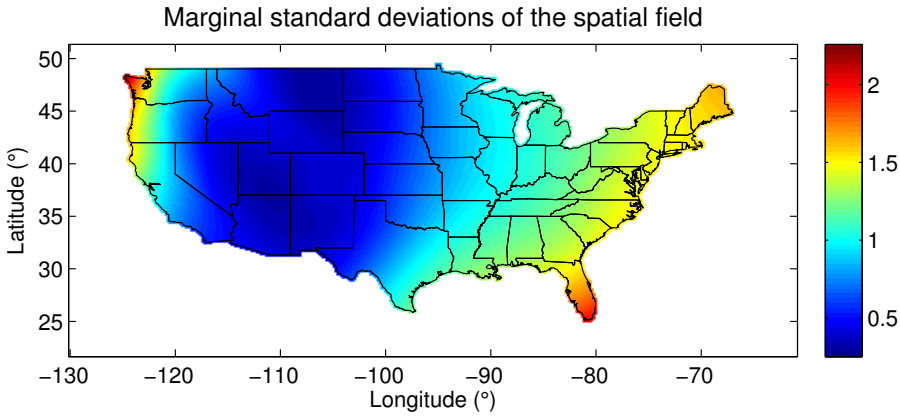
The estimated covariance structure implies strong smoothing of in the eastern region and weak smoothing in the western region. This must be understood to say something about both how well the covariates describe the data at different locations and the underlying non-stationarity in the covariance structure of the physical phenomenon. In this case there is a good fit for the elevation covariate in the mountainous areas in the western part, but it offers less information in the eastern part. From Figure 1 one can see that at around longitude 97° W there is an increase in precipitation which cannot be explained by elevation, and thus is not captured by the covariates. This jump must therefore be explained by the covariance structure, and in this case it is explained by having the covariates fit well in the western region and explaining the high values in the eastern region as being caused, randomly, by a spatial process with a long range.

In the same way as in Section 3.2 the logarithm of annual precipitation is predicted at the centre of each cell in the discretization. This gives predictions for 400×200 regularly distributed locations, where the value of the elevation covariate at each location is selected with bilinear interpolation from the closest points in the GLOBE (Hastings et al., 1999) dataset. The prediction and prediction standard deviations are shown in Figure 9. As for the stationary model, the values outside the conterminous US are coloured white. One can see that the overall look of the predictions is similar to the predictions from the stationary model, but that the prediction standard deviations differ. The prediction standard deviations vary strongly over the spatial domain because of the extreme differences in spatial range for the estimated non-stationary model.

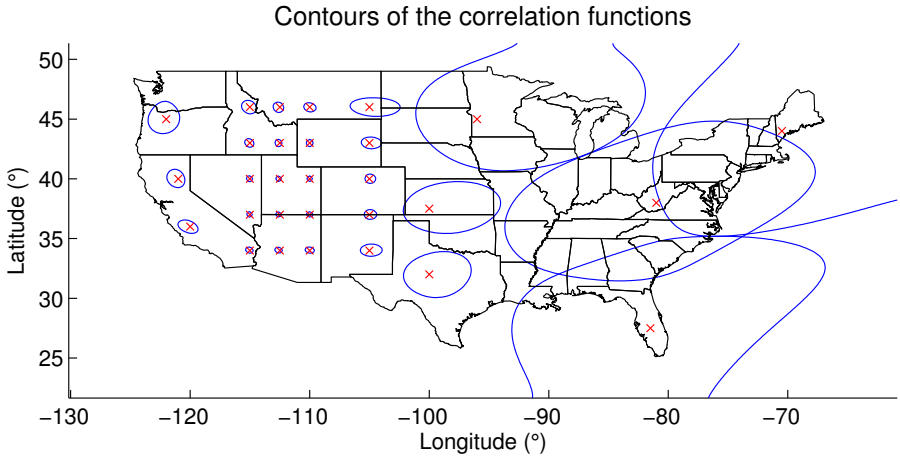
3.4 Evaluation of predictions

The predictions of the stationary model and the non-stationary model are compared with the continuous rank probability score (CRPS) (Gneiting et al., 2005) and the logarithmic scoring rule. CRPS is defined for a univariate distribution as

$$\operatorname{crps}(F, y) = \int_{-\infty}^{\infty} (F(y) - \mathbb{1}(y \leq t))^2 \, \mathrm{d}t,$$

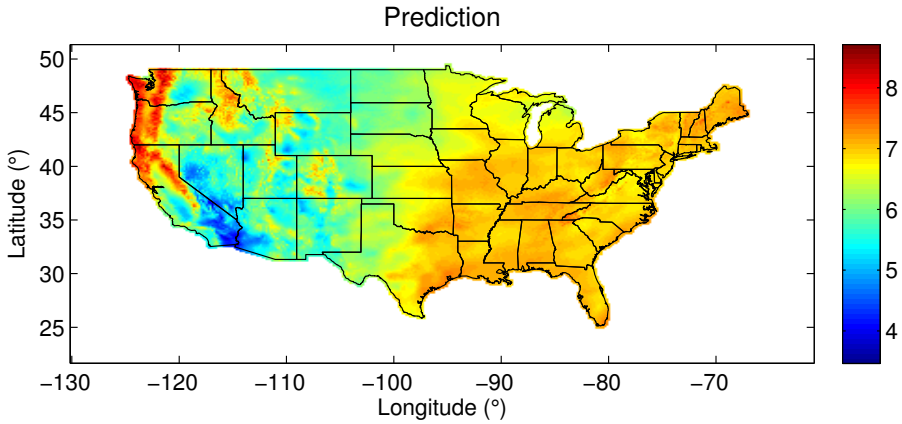


(a) Marginal standard deviations

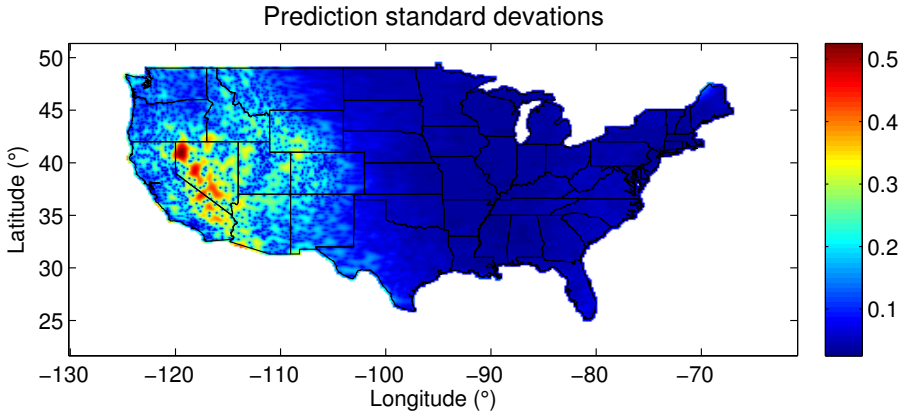


(b) Correlation structure

Figure 8: Estimated covariance structure of the spatial field. (a) Marginal standard deviations (b) Contours of 0.7 correlation for selected locations marked with red crosses



(a) Prediction



(b) Prediction standard deviations

Figure 9: Predictions and prediction standard deviations for the non-stationary model for the logarithm of annual precipitation in the conterminous US in 1981 measured in millimetres.

where F is the distribution function of interest, y is an observation and $\mathbb{1}$ is the indicator function. This gives a measure of how well a single observation fits a distribution. The total score is calculated as the average CRPS for the test data,

$$\text{CRPS} = \frac{1}{N}\sum_{i=1}^{N} \text{crps}(F_k, y_k),$$

where $\{y_k\}$ is the test data and $\{F_k\}$ are the corresponding marginal predictive distributions given the estimated covariance parameters and the training data. The logarithmic scoring rule is based on the joint predictive distribution of the test data $\vec{y}^*$ given the estimated covariance parameters $\hat{\vec{\theta}}$ and the training data $\vec{y}$,

$$\text{LogScore} = -\log \pi\left(\vec{y}^* | \hat{\vec{\theta}}, \vec{y}\right).$$

The comparison of the models is done using holdout sets where each holdout set consists of 20% of the locations chosen randomly. The remaining 80% of the locations are used to estimate the parameters and to predict the values at the locations in the holdout set. This procedure is repeated 20 times. For each repetition the CRPS, the logarithmic score and the root mean square error (RMSE) are calculated. From Figure 10 one can see that measured by both log-predictive score and CRPS the non-stationary model gives better predictions, but that the RMSE does not show any improvement.

However, the RMSE is based only on the point predictions and does not incorporate the prediction variances. The log-predictive score and the CRPS are more interesting since they say something about how well the predictive distributions fit. The difference in log-predictive score is large and indicates that the non-stationary model is better, but the difference in CRPS is small and indicates only a small improvement. The likely cause for this is that the log-predictive score evaluates the joint predictive distributions and there are difference which are not showing in the univariate predictive distributions.

The full cross-validation procedure for selecting the penalty parameters is expensive and takes weeks and must be evaluated against the potential gain in any application. The results shows that the choice of scoring rule

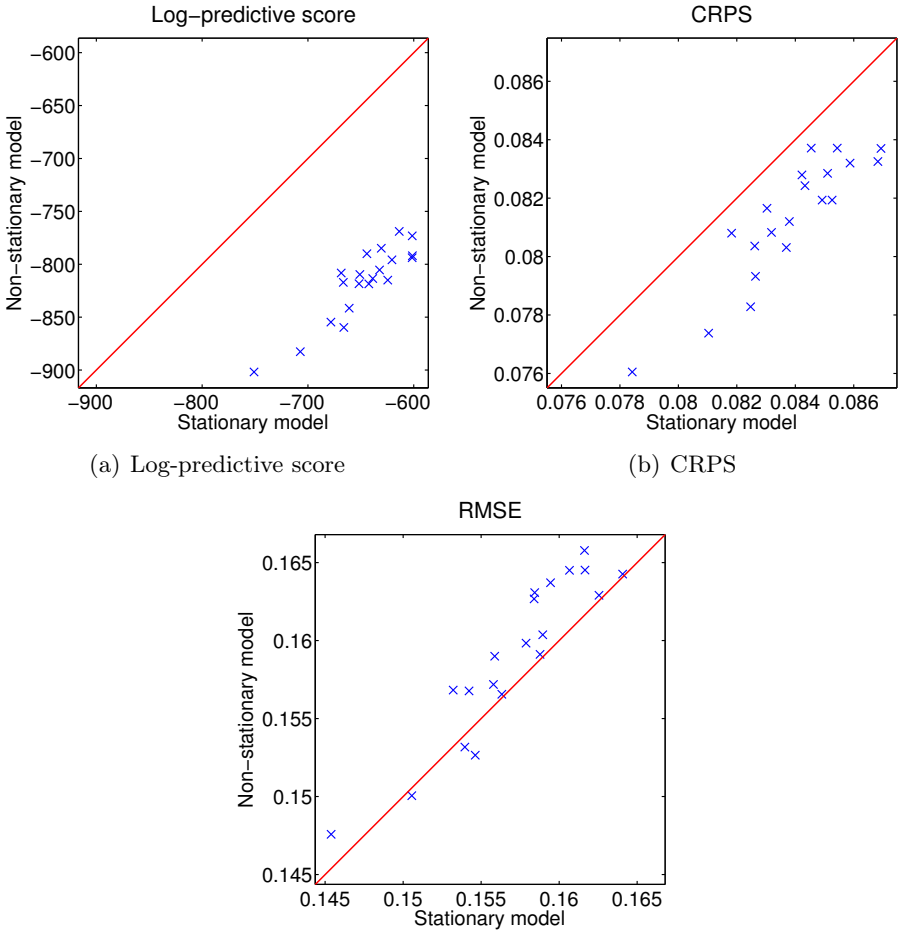
has a strong influence on the conclusion of whether the non-stationary model was worth it. The CRPS does not show evidence that all the extra computation time was worth it, but according to the log-predictive score there is a large improvement.

3.5 Criticism

The log-predictive score and CRPS are better for the non-stationary model for each hold-out set, but the covariance structure shown in Figure 8 is troubling. The range was estimated long and the marginal variances were estimated high in the eastern part because this was the "best" way to explain the changes observed, but we do not truly believe the estimates. The long estimated range means that most of the eastern part is highly correlated and the high marginal variance means that next year there might be a large change in the level in the eastern part. Whereas the low marginal variance in the west means that there will be far less changes in the spatial field there the next year. This is clearly wrong since the data for different years do not show huge changes, which are compatible with the estimated standard deviations of the spatial field, in the level of precipitation between years in the eastern region.

It is well-known that the range and the marginal variance of the stationary Matérn model are not identifiable from a fixed-size observation window (Zhang, 2004), and the situation is not likely to improve for a complex model with spatially varying marginal variances and covariance structure, but what we are seeing is the result of forcing the model to include mean structure in the covariance structure. Based on data from multiple years it is clear that the difference in level between the western and eastern region is actually caused by a change in the mean. Further, the short range in the west is also problematic because it means that few of surrounding data points are being used to predict values in this part of the domain. This could mean that the spatial effect is weak in this region, but the estimated covariance structure gives evidence that we need to investigate the cause more thoroughly.

This makes an important point regarding the worth and usefulness of the non-stationary model. Whether we have improved the CRPS and the log-predictive score is not the only question worth asking. We have gained understanding about issues in the estimated covariance structure that we



(a) Log-predictive score

(b) CRPS

(c) RMSE

Figure 10: Scatter plots of prediction scores from the stationary and the non-stationary model. 20% of the locations are randomly chosen to be held out and the remaining 80% are used to estimate parameters and predict the 20% held out data. This was repeated 20 times. For values below the line the non-stationary model is better, and conversely for values above the line.

need to investigate to understand where the non-stationarity is coming from and whether it is correctly captured in the model. In this case we have gained something more than an improvement in prediction scores. We have identified two potential issues with the model: the wrongly specified mean, which we knew about, and the weak spatial effect in the western region, which we need investigate.

4 Non-stationarity in multiple realizations

4.1 Non-stationary modelling framework

If we use multiple realizations, the non-stationarity in the mean and the non-stationarity in the covariance structure are separable. Modelling them separately goes beyond adaptive smoothing and is a situation where the term non-stationary modelling is accurate. The goal in this section is to separate out the non-stationarity in the mean and to investigate the two issues we discovered in the analysis of a single year in the previous section: over-smoothing in the eastern region and under-smoothing in the western region.

We repeat the analysis using data from the years 1971–1985, and we want to see how much the predictions improve and how the estimated non-stationary changes with a better model for the mean. Ideally, one could fit a full spatio-temporal model to these years, but since the focus is on the spatial non-stationarity we will assume that the 15 years are independent realizations of the same spatial process. Since we are using precipitation data aggregated to yearly data, the temporal dependence is weak and this is a reasonable simplification.

4.2 De-trending

The first step in the analysis is to de-trend the dataset. Each year has a different number of observations and some observations are at different locations, which means that there will be different missing locations for each year. The de-trending is done with a simple model that assumes that each year is an independent realization of a stationary spatial field and is observed with measurement noises with the same variance. The model is estimated based on the observations, and the values at locations of interest

at each year is filled in based on the posterior marginal conditional means. Then we take the average of the fitted values over the 15 year period as an estimate of the true mean.

The simple model is fitted using the R package INLA, which is based on the INLA method of (Rue et al., 2009). The model used is

$$y(\vec{s}_i, t) = \mu + x(\vec{s}_i)\beta + u_t(\vec{s}_i) + a_t + \epsilon_{i,t},\ i = 1, 2, \ldots, N_t,\ t = 1971, \ldots, 1985,$$

where μ is the joint mean for all observations, $x(\vec{s}_i)$ is the elevation at location $\vec{s}_i$ and β is the associated coefficient for the covariate, u_t for $t = 1971, 1972, \ldots, 1985$ are independent realizations of the spatial effect for each year, a_t is an AR(1) process supposed to capture temporal changes in the joint mean between years, and $\epsilon_{i,t}$ are independent Gaussian measurement errors. The spatial effect is approximately Matérn with smoothness parameter $\nu = 1$. The model is estimated and used to predict the values at all locations of interest in all 15 years. The estimate of the true mean $\hat{\mu}(\vec{s})$, at location $\vec{s}$, is found by taking the average over the estimated value at each year.

In the rest of the section we focus on the residuals $y(\vec{s}_i, t) - \hat{\mu}(\vec{s}_i)$. This means that the estimate of the mean is assumed to be without uncertainty. The intention is to remove most of the non-stationarity in the mean and then evaluate whether there is remaining non-stationarity in the covariance structure of the de-trended data that benefits from being modelled with a non-stationary model. The de-trended data for 1981 is shown in Figure 11. The de-trended data can be compared to the original data in Figure 1. One clear difference between the two figures is that the de-trended data does not have an obvious shift in the level of the precipitation between the western and eastern sides.

4.3 Fitting the non-stationary model

We fit a stationary model (STAT1) and a non-stationary model (NSTAT1) as in Section 3, but without covariates and with the assumption that there are 15 independent replications of the residuals. Each year has observations at potentially different locations, but this does not pose any problems in the SPDE-based model since the entire field is modelled explicitly through the values on each cell in the discretization. The observations are mapped to

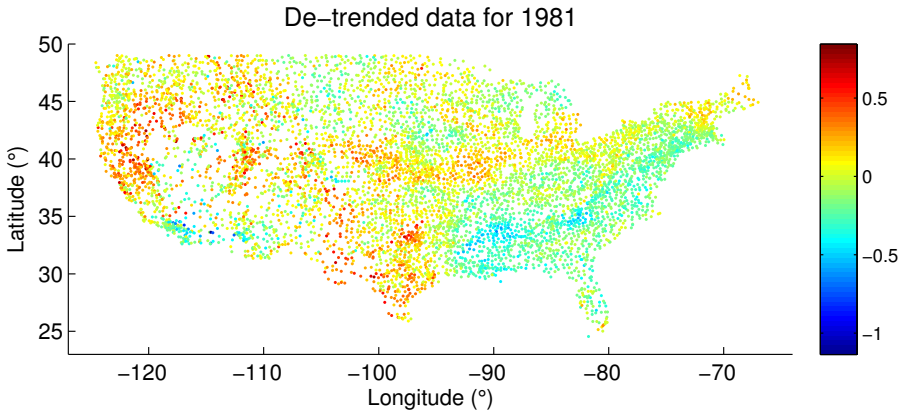


Figure 11: De-trended observations of log-transformed total annual precipitation measured in millimeter for 1981.

statements about the values on the grid cells in each year and the inference proceeds in a similar way as for the adaptive smoothing application that used only the year 1981.

The penalty parameters τ_1, τ_2, τ_3 and τ_4 should be changed, but with 15 realizations the cross-validation becomes far more computationally expensive. Therefore, we performed an exploratory analysis where the fits for low, medium and high smoothing were compared, and we decided to use $\log(\tau_1) = 10$, $\log(\tau_2) = 10$, $\log(\tau_3) = 10$ and $\log(\tau_4) = 10$. This might not lead to the highest possible decrease in the prediction scores, but at this point the main interest lies in the qualitative changes in the estimated structure. And, it would, potentially, be a waste of time to put in the required effort before we are certain that there are not major components missing in the model.

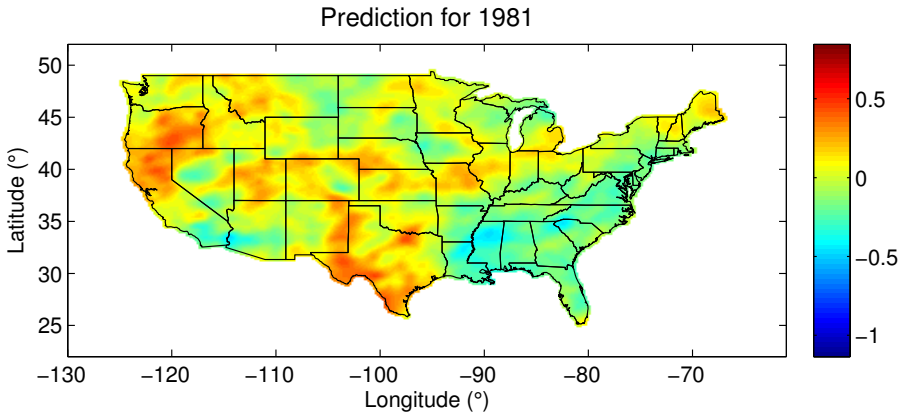
The parameters were estimated in the same way as in Section 3, and the maximum penalized likelihood estimates for non-stationarity were used to give the predictions shown in Figures 12 and 13. The figures show both the predictions and the prediction standard deviations for STAT1 and NSTAT1. There are several interesting features in these plots. First, the predicted values are similar for the two models and the main difference is found in the prediction standard deviations. Second, the prediction

standard deviations for the western region is troubling for NSTAT1. The range appears to be too short and the spatial effect appears to be close to independent measurement noise in this area. This is not consistent with Figure 11, which appears to have a spatial effect in this region as well.

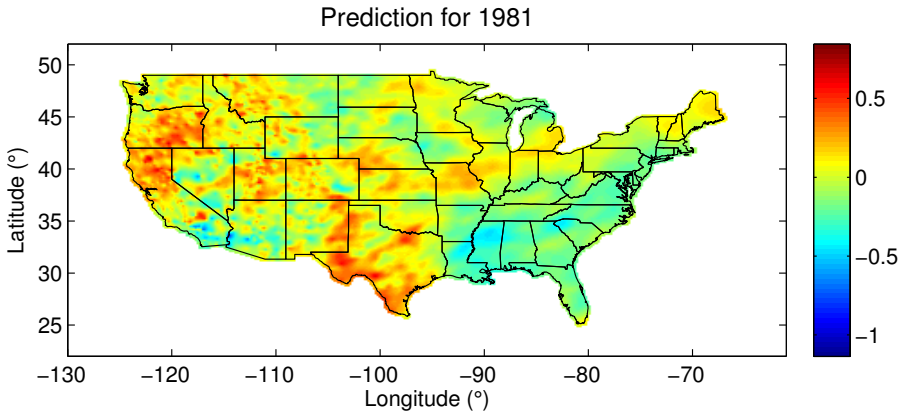
The problem can be seen clearly when looking at the estimated covariance structure shown in Figure 14. The correlation structure in the eastern part looks regular after de-trending the data, but the correlation structure in the western region is almost degenerating to independent noise. This is a problem from a computational perspective, since the discretization of the SPDE requires that the range is not too small compared to the size of the grid cells, and from a modelling perspective, since the parameters are supposed to describe a slowly changing spatial dependence structure. In the case that the spatial range is that low, the SPDE models requires a high resolution to properly capture the dependence between neighbouring grid cells in the discretization, but if the range is that low, a spatial effect might not be needed. Furthermore, Figure 14(a) shows that the variance of the spatial field is higher in the western region. This indicates that the nugget effect in the western region needs to be different from the nugget effect in the eastern region.

The fits of STAT1 and NSTAT1 are compared with the log-predictive score, the CRPS and the RMSE. The scores are calculated by randomly dividing the data in each year in five parts and then holding out the first part from each year and do the entire fitting and prediction of this data using only the remaining part of the data. Then holding out the second part of the data in each year and so on, for a total of 5 values. This process was then repeated three more times for a total of 20 values of the scores. Scatter plots comparing the scores for the two models are shown in Figure 15.

NSTAT1 has a lower log-predictive score and CRPS than STAT1, but the RMSE is higher. The conclusions based on the log-predictive score and the CRPS is the same as for the single realization analysis in Section 3.4. However, the consistently higher RMSE values indicate that there is a problem with the model. The problem lies in the western region where the range is too low, which leads to worse point estimates because the spatial dependence is not exploited. The flexible non-stationary model is able to detect that a higher variance is required for the nugget effect in the western

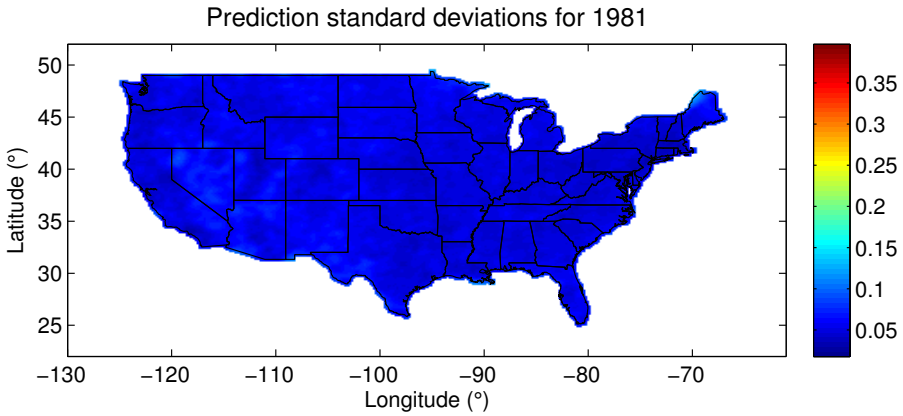


(a) Prediction for STAT1

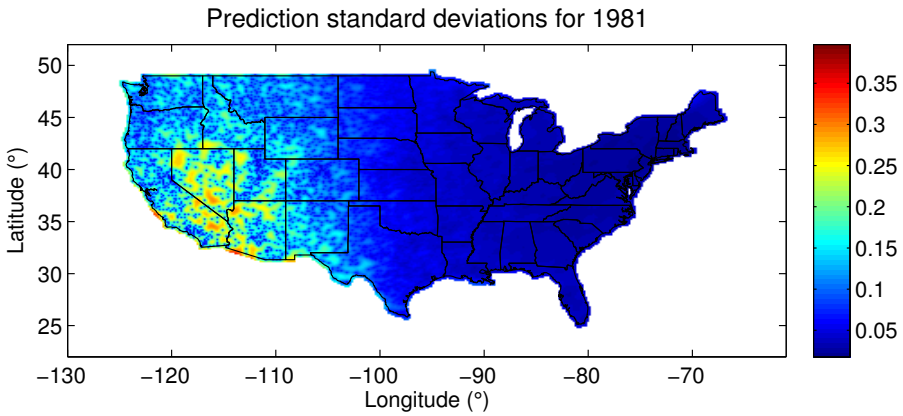


(b) Prediction for NSTAT1

Figure 12: Prediction for de-trended data for year 1981 based on the 15 year period 1971–1985. (a) shows the prediction for STAT1, (b) shows the prediction for NSTAT1.

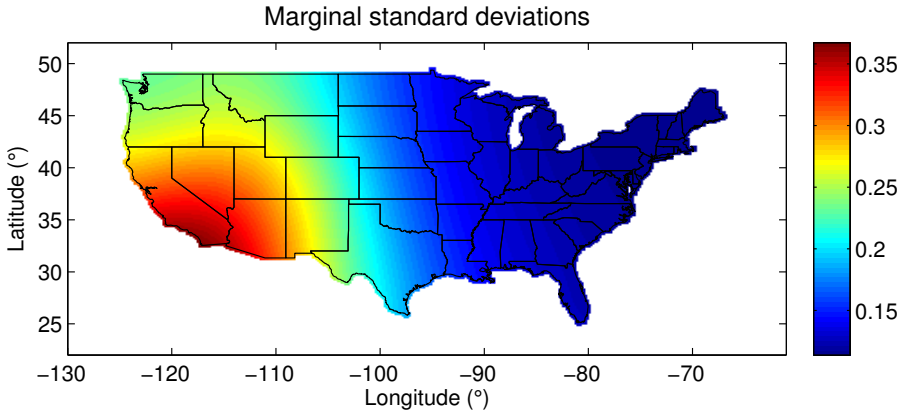


(a) Prediction standard deviations for STAT1

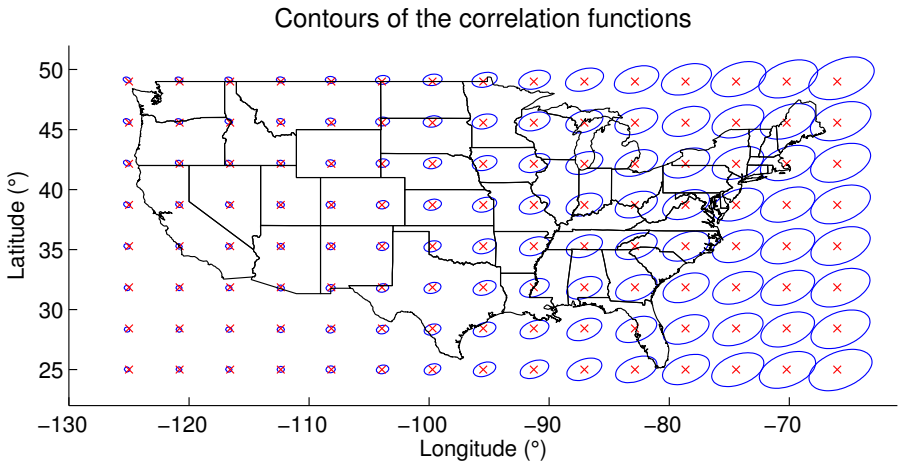


(b) Prediction standard deviations for NSTAT1

Figure 13: Prediction for de-trended data for year 1981 based on the 15 year period 1971–1985. (a) shows the prediction standard deviations for STAT1 and (b) shows the prediction standard deviations for NSTAT1.



(a) Marginal standard deviations



(b) Estimated correlation structure

Figure 14: (a) Estimated marginal standard deviations and (b) estimated 0.7 level contour curves for the correlation functions with respect to the locations marked with red crosses for the spatial effect in NSTAT1.

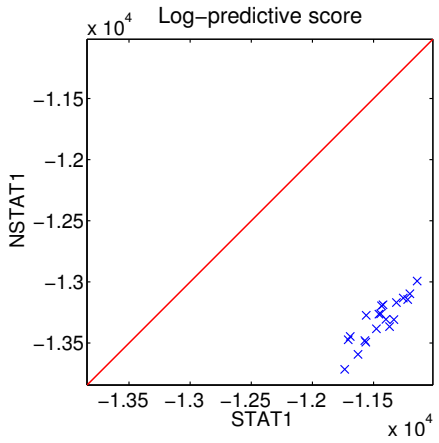
region, but is not able to achieve this in the correct way. Even with all the freedom available in the model it is impossible to have spatial dependence and different nugget effects because we have put the non-stationarity in the wrong components of the model. We need to treat the nugget effects in the western and eastern regions separately.

4.4 Removing the under-smoothing in the western part

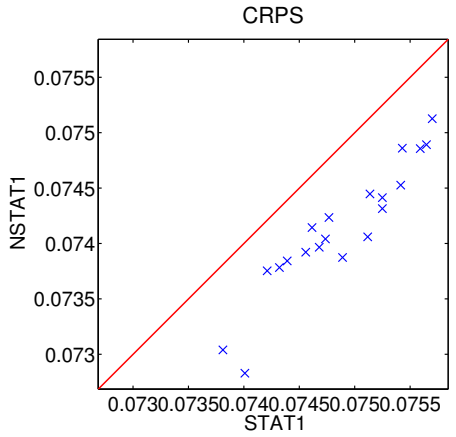
The results in Section 4.3 indicate that the nugget effect is different in the western and the eastern part of the conterminous US. Therefore, we fit a stationary model (STAT2) and a non-stationary model (NSTAT2) with separate nugget effects for locations with longitudes lower than $100\,°W$ and for locations with longitudes higher than or equal to $100\,°W$. The placement of the frontier at $100\,°W$ is motivated by the change from mountainous regions to plains seen in Figure 2 and the change from low to high range seen in Figure 14(b), but we do not believe it would be particularly sensitive to the exact placement as long as it is in the area of transition from mountainous regions to plains. Except for this change, the models are unchanged, and we use the same penalties τ_1, τ_2, τ_3 and τ_4 for the non-stationarity structure. The intention is to see how much the predictions and the estimated dependence structure change with different nugget effects, but the same penalties.

The predictions and prediction standard deviations are shown in Figures 16 and 17, respectively. Figure 17 show that the prediction standard deviations for NSTAT2 do not have the strange artifacts in the western region that are present in Figure 13 for NSTAT1, but one can notice that there is a sharp change in prediction standard deviations at longitude $100\,°W$. This is by construction due to the use of different nugget effects for the two parts of the conterminous US. STAT2 has an estimated standard deviation for the nugget effect of 0.17 in the western part and of 0.083 in the eastern part and for NSTAT2 the estimated standard deviation for the nugget effect is 0.16 in the western part and is 0.083 in the eastern part.

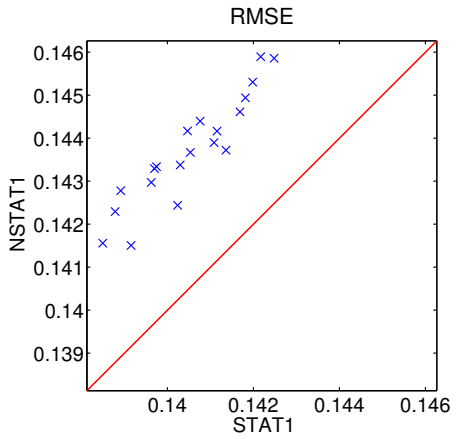
The estimated spatial dependence structure of NSTAT2 is shown in Figure 18. The clearest change from the dependence structure of NSTAT1 shown in Figure 14 is that the non-stationarity in the correlation structure is mostly gone. The appearance is much more reasonable than for NSTAT1



(a) Log-predictive score

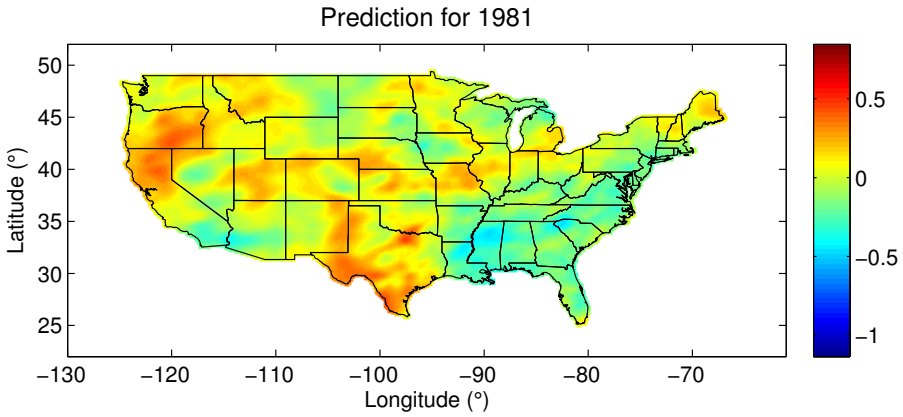


(b) CRPS

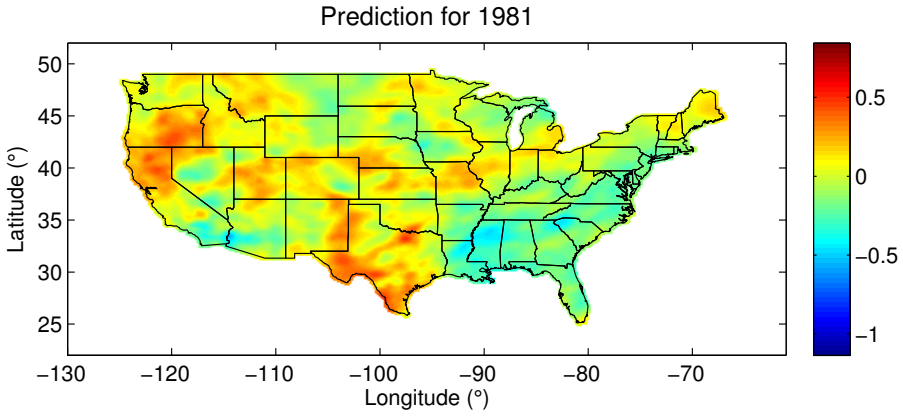


(c) RMSE

Figure 15: Scatter plots of (a) Log-predictive score, (b) CRPS and (c) Root mean square error for STAT1 and NSTAT1. The estimates were calculated with hold-out sets where 20% of the locations were held-out from each year as described in Section 4.3.

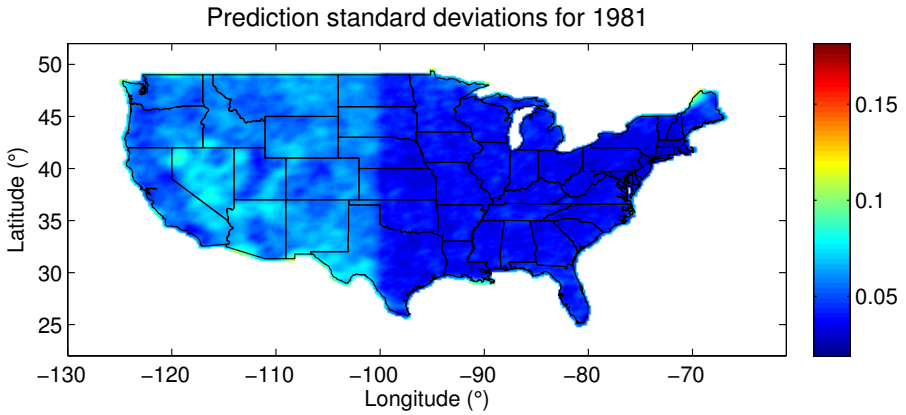


(a) Prediction for STAT2

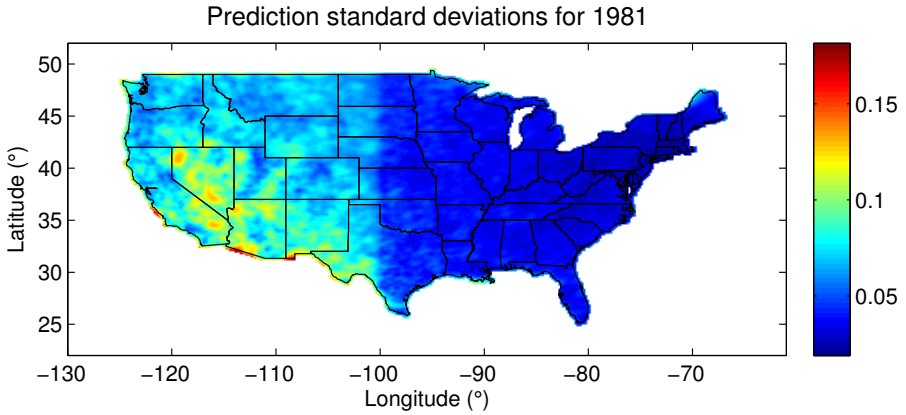


(b) Prediction for NSTAT2

Figure 16: Prediction for de-trended data for year 1981 based on the 15 year period 1971–1985. (a) shows the prediction for STAT2, (b) shows the prediction for NSTAT2.



(a) Prediction standard deviations for STAT2



(b) Prediction standard deviations for NSTAT2

Figure 17: Prediction for de-trended data for year 1981 based on the 15 year period 1971–1985. (a) shows the prediction standard deviations for STAT2 and (b) shows the prediction standard deviations for NSTAT2.

since the entire dependence structure is changing slowly and there are no areas with unreasonably large or small ranges. Some non-stationarity still remains in the marginal standard deviations, but together these plots indicate that the simple model STAT2, which does not use a complex non-stationary spatial field, should fit these data well.

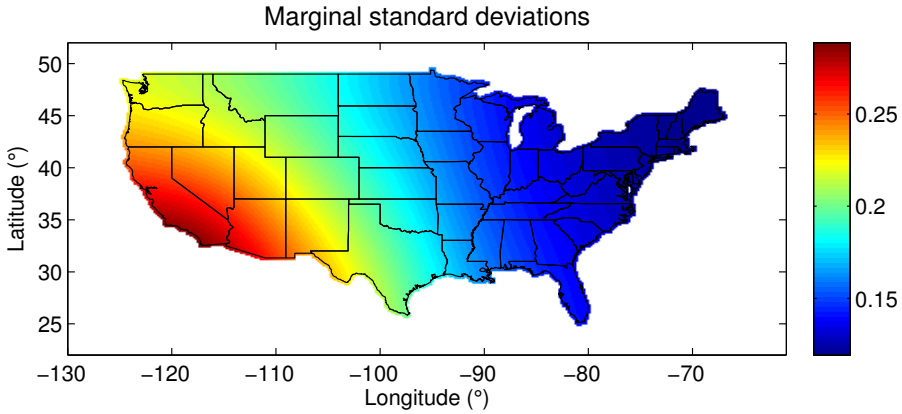
We compare the predictions of STAT2 and NSTAT2 by the RMSE, the CRPS and the log-predictive score. The results are given in Figure 19. NSTAT2 performs better according to all of the scores. The scatter plots of the scores show that NSTAT2 performs better for all the hold-out sets, but that the differences in scores are small.

4.5 Discussion of models

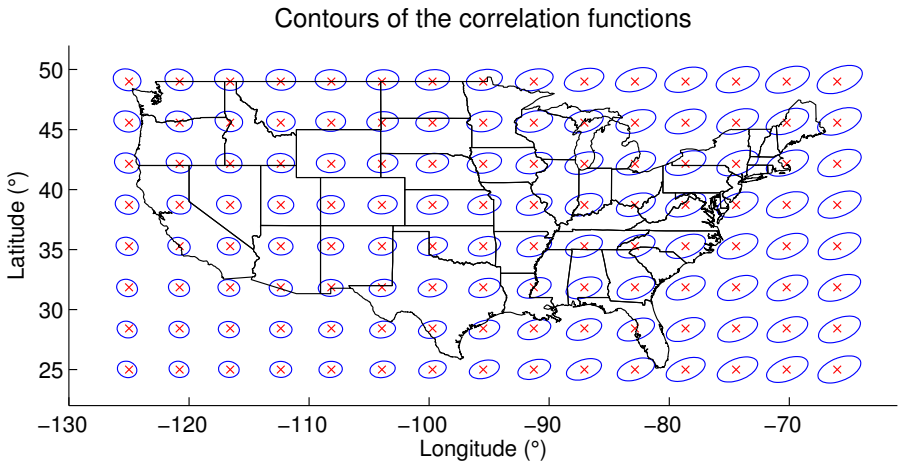
The prediction scores for STAT1, NSTAT1, STAT2 and NSTAT2 are shown in Figure 20. The figure shows that the model performing the best according to all scores is NSTAT2, but is the extra computation time worth the effort in this case? The much simpler model STAT2 is performing almost as good as NSTAT2 and requires only *one* extra parameter. The cost of including one extra parameter is far less than the cost of introducing the flexible non-stationary model. Additionally, one can see that even though the expensive flexible model makes NSTAT1 consistently better than STAT1 in the log-predictive score and the CRPS, STAT2 makes an even greater improvement from STAT1 for the cost of only a single parameter.

The predictions and prediction standard deviations for STAT2 and NSTAT2 in Figures 16 and 17 are showing less extreme differences than the predictions and prediction standard deviations for STAT1 and NSTAT1 shown in Figures 12 and 13, but there is still some differences in the prediction standard deviations. Some further gain is possible by selecting the penalty parameters controlling the non-stationarity more carefully. We saw some improvement by trying different penalty parameters, but no major changes that would change the conclusion. When we take computation time into account, STAT2 appears to be the better choice. There is some gain with the flexible non-stationary model in NSTAT2, but it comes at a high computational cost.

The physical cause of the difference in the nugget effect between the western region and the eastern region is not known, but it is unlikely

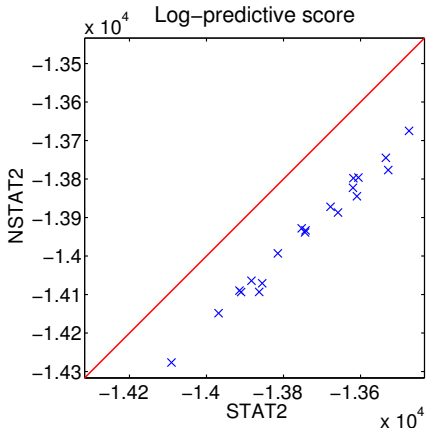


(a) Marginal standard deviations

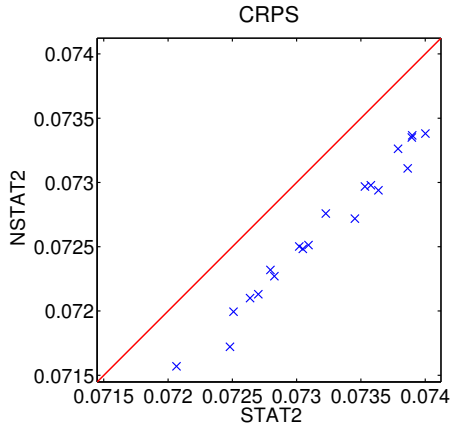


(b) Estimated correlation structure

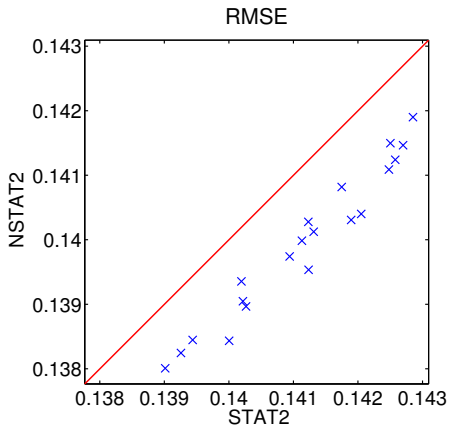
Figure 18: (a) Estimated marginal standard deviations and (b) estimated 0.7 level contour curves for the correlation functions with respect to the locations marked with red crosses for the spatial effect in NSTAT2.



(a) Log-predictive score

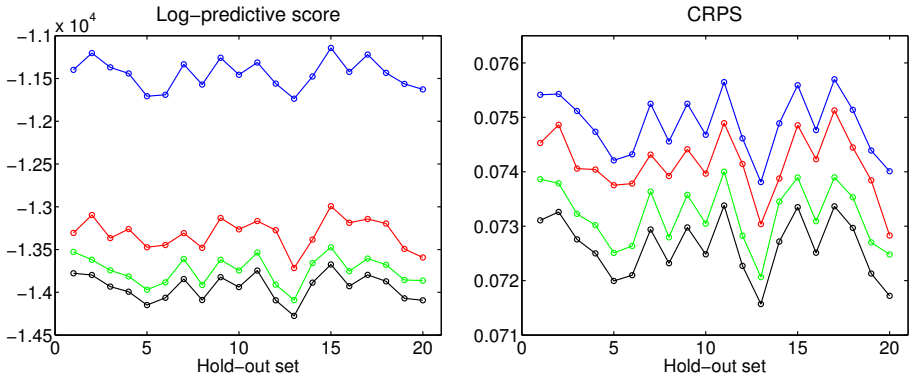


(b) CRPS



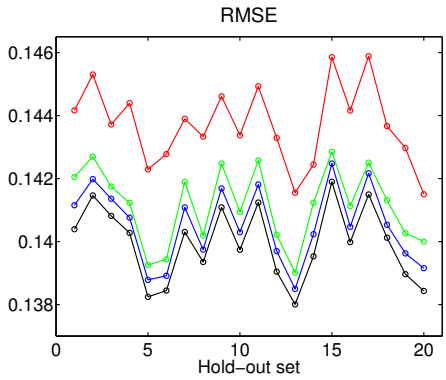
(c) RMSE

Figure 19: Scatter plots of (a) log-predictive score, (b) CRPS and (c) Root mean square error for STAT2 and NSTAT2. The estimates were calculated with hold-out sets where 20% of the locations were held-out from each year as described in Section 4.3.



(a) LPD

(b) CRPS



(c) RMSE

Figure 20: Comparison of STAT1 (blue), NSTAT1 (red), STAT2 (green) and NSTAT2 (black) based on (a) Log-predictive score, (b) CRPS and (c) RMSE. The estimates were calculated with hold-out sets where 20% of the locations were held-out from each year as described in Section 4.3.

to be caused only by differences in the measurement equipment. It is more likely that it is caused by differences in the small-scale behaviour of the process generating the weather in the two different regions that is not captured by the model, but it has not been our intention to find the physical explanation. The intention has been to demonstrate how such a phenomenon can affect the estimation of general flexible models for non-stationarity and the need to carefully evaluate the fitted covariance structures.

5 Discussion

The question of whether we need non-stationary spatial models or not, is a deeper question than it might seem initially. The first step of the analysis should be to decide whether it is likely that non-stationarity is present in the data or not, and in this context simple data exploration, such as variograms, and formal tests (Fuentes, 2005; Jun and Genton, 2012; Bowman and Crujeiras, 2013) are useful tools. The second step is to decide which non-stationary model we want to use and it can be tempting to look for complex models that allow for spatial fields that have large amounts of flexibility in the covariance structure. We then apply these models with the hope that the high degree of flexibility means that we will be able to capture any non-stationarity present in the data, but the analysis of the annual precipitation data shows that blindly applying such a model might not capture the non-stationarity in the correct and best way.

The case study clearly indicates the need to go beyond simply determining whether or not non-stationarity is present in the data. We need to determine what type of non-stationarity that is present in the data. A flexible model will try to adapt to the non-stationarity, but if the flexibility is available in the wrong parts of the model, the model might have to do suboptimal things to improve the predictive distributions. For example, imitate a spatially varying nugget effect by decreasing the range and varying the marginal variances. This adaptation gives severe undersmoothing, but simply expanding the model with a smoothly varying nugget effect would make the model difficult to identify together with the rest of the flexibility. Therefore, we should determine what is causing the non-stationarity

we are seeing before deciding which non-stationary model to use.

The first and most obvious source of non-stationarity in a dataset is the mean structure, and not accounting for this source of non-stationarity will confound the non-stationarity in the mean structure with the non-stationarity in the covariance structure. For example, unmeasured covariates can lead to the apparent long range dependence and global non-stationarity that we observed in the analysis of a single realization. The method presented in this paper is aimed at modelling local non-stationarity and is not appropriate for modelling this type of global non-stationarity. We handle this apparent structure in the covariances by de-trending the data, but it is also possible to model jointly the mean structure and the covariance structure. A simple example of the latter would be to combine the SPDE models with a small number of global basis functions to form a hybrid of fixed-rank kriging and the SPDE models, where the SPDE models captures the short range dependence and local non-stationarity, and the basis functions capture the long range dependence and global non-stationarity. Whichever approach is taken, the paper demonstrates the need to remove the global non-stationarity before modelling the local non-stationarity.

After we have removed the global non-stationarity induced by the mean structure we can model the remaining local non-stationarity, for which the Markovian structure of the SPDE models offers a good modelling tool. In the SPDE models we construct a consistent global covariance structure by tying together the local behaviour specified by the SPDE at each location, and the covariance between any two locations will be a combination of the local behaviour at all locations in the model. We believe that this approach is a good way to model local non-stationarity that provides a more flexible, more computationally efficient and easier to parametrize approach than the deformation method, while still having a geometric interpretation of varying the local distance measures.

But modelling local non-stationarity requires information on the small-scale directional behaviour of the observations, and we would be hesitant to estimate flexible non-stationary models for sparser datasets. Methods such as the deformation method is routinely applied to much sparser datasets, but there is no way around the fact that for patches where we do not have observations we have no idea how the covariances behave. For sparse data

it is possible to imagine multiple covariance structures that could give rise to the observed empirical covariances and the unobserved structure must be filled by the model based on the assumptions and restrictions that we have put into the model. This can, potentially, lead to highly model dependent estimates since in non-stationary modelling the missing covariances do not directly affect the observations, and it is important to not allow too much freedom in the covariance structure compared to the sparseness of the data, and to realize that the features seen in the estimated covariance structure will depend on the sparseness of the data.

In an analogous way as for other finite-dimensional methods, there is a confounding of the nugget effect and the resolution chosen for the finite-dimensional approximation. For predictive processes there exists a solution (Finley et al., 2009), but for the SPDE models it is an active field of research. In a GRF model the nugget effect is a combination of the small-scale behaviour and the measurement error, where small-scale behaviour is behaviour below the scale which the data can inform about. The sparser the data is, the more small-scale variation will be confounded with the nugget effect, but for the SPDE models the interpretation of the nugget effect is also tied to the discretization and is a combination of measurement error, small-scale variation and sub-grid variation. The approximation cannot capture variation within the grid cells and these variations increase the nugget variance and decrease the process variances, but this is only a worry when interpreting these parameters. If the precipitation data were sparser, the confounding between small-scale variation and the nugget effect would make it difficult to detect different nugget effects in the western region and the eastern region, and the approach might lead to a different conclusion about the nugget effect.

In each of the three cases studied, the flexible non-stationary model performs better according to the log-predictive score and the CRPS, but when we target directly the non-stationarity in the nugget effect, we can apply a much simpler model just using two nugget effects. Does this mean that the flexible non-stationary model was not useful? No, we were able to use the flexible non-stationary model to estimate a covariance structure that could be used to help determine possible sources of the non-stationarity. We could then include these sources directly and fit a simpler model performing almost equally well, and we could make the same changes to the

flexible non-stationary model and fit it again to become confident that there were no other major uncaptured sources of non-stationarity. The idea that the nugget might be the source of heterogeneity is not new (Zimmerman, 1993), but the case study demonstrates the dangers of putting the heterogeneity in the wrong components in the model.

If there were knowledge available about what was physically generating the non-stationarity, it would be possible to make simpler models where we reduce the flexibility and control the the covariance structure by covariates. The use of two nugget effects is an extreme case of this, but covariates in the covariance structure has been a recent direction of research within all the major families of approaches such as the deformation method, the process convolution method and the SPDE-based method (Schmidt et al., 2011; Neto et al., 2014; Ingebrigtsen et al., 2014). However, even if we intend to use covariates, the more general non-stationary models could be used to gain intuition about which covariates should be selected and what type of non-stationarity they should control.

The comparison of the different models shows that the scoring rule used to evaluate the predictions has a large influence on the conclusion. The use of a non-stationary model instead of a stationary model mainly affects the prediction variances and not the predicted values. Therefore, the largest improvements are seen in the log-predictive score and the CRPS, and not the RMSE that only evaluates point predictions. However, consistently higher RMSE values for the flexible non-stationary model compared to the simple stationary, as observed when fitting the models using a single nugget effect to de-trended data, is useful to detect problems with the model such as undersmoothing.

One of the major reasons not to use general non-stationary models unless they are absolutely needed is that they are computationally expensive. The covariate-based approach is less expensive, but requires assumptions about how the non-stationarity varies. Another approach would be to estimate the model locally in different parts of the domain and then try to piece everything together for predictions, but looking for the most efficient way to estimate the model is not the goal of this paper and the more complex one makes the model, the more computationally expensive it will be. The point we are trying to make is that in applications, time might in many cases be better spent on considering how to put the non-stationarity

into the model than on developing more complex flexible models and ways to compute them.

Non-stationarity in the covariance structure of spatial models is needed even after the non-stationarity in the mean has been removed, but we need to think carefully about how we handle the non-stationarity. We need to go beyond determining whether there is non-stationarity or not, and determine what type of non-stationarity is present and if possible target this non-stationarity directly instead of using a general flexible model. But in this context the estimated covariance structure from a general flexible model can in some cases be a useful tool to determine how to do this.

A Derivation of the second-order random walk prior

Each function, f, is a priori modelled as a Gaussian process described by the SPDE

$$-\Delta f(\vec{s}) = \frac{1}{\sqrt{\tau}}\mathcal{W}(\vec{s}), \qquad \vec{s} \in \mathcal{D} = [A_1, B_1] \times [A_2, B_2], \tag{A.1}$$

where $A_1 < B_1$, $A_2 < B_2$ and $\tau > 0$, $\mathcal{W}$ is standard Gaussian white noise and $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$, with the Neumann boundary condition of zero normal derivatives at the edges. In practice this is approximated by representing f as a linear combination of basis elements $\{f_{ij}\}$ weighted by Gaussian distributed weights $\{\alpha_{ij}\}$,

$$f(\vec{s}) = \sum_{i=1}^{K}\sum_{j=1}^{L} \alpha_{ij} f_{ij}(\vec{s}).$$

The basis functions are constructed from separate bases $\{g_i\}$ and $\{h_j\}$ for the x-coordinate and the y-coordinate, respectively,

$$f_{ij}(\vec{s}) = g_i(x)h_j(y). \tag{A.2}$$

For convenience each basis function is assumed to fulfill the boundary condition of zero normal derivative at the edges.

Let $\vec{\alpha} = \mathrm{vec}([\alpha_{ij}]_{ij})$, then the task is to find the best Gaussian distribution for $\vec{\alpha}$. Where "best" is used in the sense of making the resulting distribution for f "close" to a solution of SPDE (A.1). This is done by a least-squares approach where the vector created from doing inner products of the left hand side with $-\Delta f_{kl}$ must be equal in distribution to the vector created from doing the same to the right hand side,

$$\mathrm{vec}\left([\langle -\Delta f, -\Delta f_{kl}\rangle_{\mathcal{D}}]_{kl}\right) \overset{d}{=} \mathrm{vec}\left([\langle \mathcal{W}, -\Delta f_{kl}\rangle_{\mathcal{D}}]_{kl}\right). \tag{A.3}$$

First, calculate the inner product that is needed

$$\begin{aligned}\langle -\Delta g_i h_j, -\Delta g_k h_l\rangle_{\mathcal{D}} &= \langle \Delta g_i h_j, \Delta g_i h_j\rangle_{\mathcal{D}} \\ &= \left\langle \left(\frac{\partial^2}{\partial x^2} g_i\right) h_j + g_i \frac{\partial^2}{\partial y^2} h_j, \left(\frac{\partial^2}{\partial x^2} g_k\right) h_l + g_k \frac{\partial^2}{\partial y^2} h_l \right\rangle_{\mathcal{D}}.\end{aligned}$$

The bilinearity of the inner product can be used to expand the expression in a sum of four innerproducts. Each of these inner products can then be written as a product of two inner products. Due to lack of space this is not done explicitly, but one of these terms is, for example,

$$\left\langle \left(\frac{\partial^2}{\partial x^2} g_i\right) h_j, \left(\frac{\partial^2}{\partial x^2} g_k\right) h_l \right\rangle_{\mathcal{D}} = \left\langle \frac{\partial^2}{\partial x^2} g_i, \frac{\partial^2}{\partial x^2} g_k \right\rangle_{[A_1,B_1]} \langle h_j, h_l \rangle_{[A_2,B_2]}.$$

By inserting Equation (A.2) into Equation (A.3) and using the above derivations together with integration by parts one can see that the left hand side becomes

$$\text{vec}\left([\langle -\Delta f, -\Delta f_{kl}\rangle_{\mathcal{D}}]_{kl}\right) = \mathbf{C}\vec{\alpha},$$

where $\mathbf{C} = \mathbf{G}_2 \otimes \mathbf{H}_0 + 2\mathbf{G}_1 \otimes \mathbf{H}_1 + \mathbf{G}_0 \otimes \mathbf{H}_2$ with

$$\mathbf{G}_n = \left[\left\langle \frac{\partial^n}{\partial x^n} g_i, \frac{\partial^n}{\partial x^n} g_j \right\rangle_{[A_1,B_1]} \right]_{i,j}$$

and

$$\mathbf{H}_n = \left[\left\langle \frac{\partial^n}{\partial y^n} h_i, \frac{\partial^n}{\partial y^n} h_j \right\rangle_{[A_2,B_2]} \right]_{i,j}.$$

The right hand side is a Gaussian random vector where the covariance between the position corresponding to α_{ij} and the position corresponding to α_{kl} is given by

$$\langle -\Delta f_{ij}, -\Delta f_{kl}\rangle_{\mathcal{D}}.$$

Thus the covariance matrix of the right hand side must be $\mathbf{C}$ and Equation (A.3) can be written in matrix form as

$$\mathbf{C}\vec{\alpha} = \mathbf{C}^{1/2}\vec{z},$$

where $\vec{z} \sim \mathcal{N}_{KL}(\vec{0}, \mathbf{I}_{KL})$. This means that $\vec{\alpha}$ should be given the precision matrix $\mathbf{Q} = \mathbf{C}$. Note that $\mathbf{C}$ might be singular due to invariance to some linear combination of the basis elements.

B Conditional distributions

From the hierarchical model

$$\begin{aligned}&\text{Stage 1: } \vec{y}|\vec{z},\vec{\theta} \sim \mathcal{N}_N(\mathbf{S}\vec{z}, \mathbf{I}_N/\tau_{\text{noise}})\\ &\text{Stage 2: } \vec{z}|\vec{\theta} \sim \mathcal{N}_{mn+p}(\vec{0}, \mathbf{Q}_z^{-1}),\end{aligned}$$

the posterior distribution $\pi(\vec{\theta}|\vec{y})$ can be derived explicitly. There are three steps involved.

B.1 Step 1

Calculate the distribution $\pi(\vec{z}|\vec{\theta},\vec{y})$ up to a constant,

$$\begin{aligned}\pi(\vec{z}|\vec{\theta},\vec{y}) &\propto \pi(\vec{z},\vec{\theta},\vec{y})\\ &= \pi(\vec{\theta})\pi(\vec{z}|\vec{\theta})\pi(\vec{y}|\vec{\theta},\vec{z})\\ &\propto \exp\left(-\frac{1}{2}(\vec{z}-\vec{0})^{\mathrm{T}}\mathbf{Q}_z(\vec{z}-\vec{0}) - \frac{1}{2}(\vec{y}-\mathbf{S}\vec{z})^{\mathrm{T}}\mathbf{I}_N\cdot\tau_{\text{noise}}(\vec{y}-\mathbf{S}\vec{z})\right)\\ &\propto \exp\left(-\frac{1}{2}\left(\vec{z}^{\mathrm{T}}(\mathbf{Q}_z+\tau_{\text{noise}}\mathbf{S}^{\mathrm{T}}\mathbf{S})\vec{z} - 2\vec{z}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\vec{y}\cdot\tau_{\text{noise}}\right)\right)\\ &\propto \exp\left(-\frac{1}{2}(\vec{z}-\vec{\mu}_{\mathrm{C}})^{\mathrm{T}}\mathbf{Q}_{\mathrm{C}}(\vec{z}-\vec{\mu}_{\mathrm{C}})\right),\end{aligned}$$

where $\mathbf{Q}_{\mathrm{C}} = \mathbf{Q}_z + \mathbf{S}^{\mathrm{T}}\mathbf{S}\cdot\tau_{\text{noise}}$ and $\mu_{\mathrm{C}} = \mathbf{Q}_{\mathrm{C}}^{-1}\mathbf{S}^{\mathrm{T}}\vec{y}\cdot\tau_{\text{noise}}$. This is recognized as a Gaussian distribution

$$\vec{z}|\vec{\theta},\vec{y} \sim \mathcal{N}_N(\vec{\mu}_{\mathrm{C}}, \mathbf{Q}_{\mathrm{C}}^{-1}).$$

B.2 Step 2

Integrate out $\vec{z}$ from the joint distribution of $\vec{z}$, $\vec{\theta}$ and $\vec{y}$ via the Bayesian rule,

$$\begin{aligned}\pi(\vec{\theta},\vec{y}) &= \frac{\pi(\vec{\theta},\vec{z},\vec{y})}{\pi(\vec{z}|\vec{\theta},\vec{y})}\\ &= \frac{\pi(\vec{\theta})\pi(\vec{z}|\vec{\theta})\pi(\vec{y}|\vec{z},\vec{\theta})}{\pi(\vec{z}|\vec{\theta},\vec{y})}.\end{aligned}$$

The left hand side of the expression does not depend on the value of $\vec{z}$, therefore the right hand side may be evaluated at any desired value of $\vec{z}$. Evaluating at $\vec{z} = \vec{\mu}_{\mathrm{C}}$ gives

$$\begin{aligned}\pi(\vec{\theta}, \vec{y}) &\propto \frac{\pi(\vec{\theta})\pi(\vec{z} = \vec{\mu}_{\mathrm{C}})\pi(\vec{y}|\vec{z} = \vec{\mu}_{\mathrm{C}}, \vec{\theta})}{\pi(\vec{z} = \vec{\mu}_{\mathrm{C}}|\vec{\theta}, \vec{y})} \\ &\propto \pi(\vec{\theta}) \frac{|\mathbf{Q}_z|^{1/2}|\mathbf{I}_N \cdot \tau_{\text{noise}}|^{1/2}}{|\mathbf{Q}_{\mathrm{C}}|^{1/2}} \exp\left(-\frac{1}{2}\vec{\mu}_{\mathrm{C}}^{\mathrm{T}}\mathbf{Q}_z\vec{\mu}_{\mathrm{C}}\right) \times \\ &\times \exp\left(-\frac{1}{2}(\vec{y} - \mathbf{S}\vec{\mu}_{\mathrm{C}})^{\mathrm{T}}\mathbf{I}_N \cdot \tau_{\text{noise}}(\vec{y} - \mathbf{S}\vec{\mu}_{\mathrm{C}})\right) \times \\ &\times \exp\left(+\frac{1}{2}(\vec{\mu}_{\mathrm{C}} - \vec{\mu}_{\mathrm{C}})^{\mathrm{T}}\mathbf{Q}_{\mathrm{C}}(\vec{\mu}_{\mathrm{C}} - \vec{\mu}_{\mathrm{C}})\right).\end{aligned}$$

B.3 Step 3

Condition on $\vec{y}$ to get the desired conditional distribution,

$$\begin{aligned}\log(\pi(\vec{\theta}|\vec{y})) = \text{Const} + \log(\pi(\vec{\theta})) + \frac{1}{2}\log(\det(\mathbf{Q}_z)) + \frac{N}{2}\log(\tau_{\text{noise}}) + \\ - \frac{1}{2}\log(\det(\mathbf{Q}_{\mathrm{C}})) - \frac{1}{2}\vec{\mu}_{\mathrm{C}}^{\mathrm{T}}\mathbf{Q}_z\vec{\mu}_z - \frac{\tau_{\text{noise}}}{2}(\vec{y} - \mathbf{S}\vec{\mu}_{\mathrm{C}})^{\mathrm{T}}(\vec{y} - \mathbf{S}\vec{\mu}_{\mathrm{C}}).\end{aligned} \tag{B.1}$$

C Analytic expression for the gradient

This appendix shows the derivation of the derivative of the log-likelihood. Choose the evaluation point $\vec{z} = \vec{0}$ in Appendix B.2 to find

$$\begin{aligned}&\log(\pi(\vec{\theta}, \tau_{\text{noise}}|\vec{y})) \\ &\quad = \text{Const} + \log(\pi(\vec{\theta}, \tau_{\text{noise}})) + \frac{1}{2}\log(\det(\mathbf{Q}_z)) + \\ &\quad + \frac{N}{2}\log(\tau_{\text{noise}}) - \frac{1}{2}\log(\det(\mathbf{Q}_{\mathrm{C}})) - \frac{\tau_{\text{noise}}}{2}\vec{y}^{\mathrm{T}}\vec{y} + \frac{1}{2}\vec{\mu}_{\mathrm{C}}^{\mathrm{T}}\mathbf{Q}_{\mathrm{C}}\vec{\mu}_{\mathrm{C}}.\end{aligned}$$

This is just a rewritten form of Equation (B.1) which is more convenient for the calculation of the gradient, and which separates the τ_{noise} parameter from the rest of the covariance parameters. First some preliminary results

are presented, then the derivatives are calculated with respect to θ_i and lastly the derivatives are calculated with respect to $\log(\tau_{\text{noise}})$.

Begin with simple preliminary formulas for the derivatives of the conditional precision matrix with respect to each of the parameters,

$$\frac{\partial}{\partial \theta_i}\mathbf{Q}_{\text{C}} = \frac{\partial}{\partial \theta_i}(\mathbf{Q} + \mathbf{S}^{\text{T}}\mathbf{S} \cdot \tau_{\text{noise}}) = \frac{\partial}{\partial \theta_i}\mathbf{Q} \tag{C.1}$$

and

$$\frac{\partial}{\partial \log(\tau_{\text{noise}})}\mathbf{Q}_{\text{C}} = \frac{\partial}{\partial \log(\tau_{\text{noise}})}(\mathbf{Q} + \mathbf{S}^{\text{T}}\mathbf{S} \cdot \tau_{\text{noise}}) = \mathbf{S}^{\text{T}}\mathbf{S} \cdot \tau_{\text{noise}}. \tag{C.2}$$

C.1 Derivative with respect to θ_i

First the derivatives of the log-determinants can be handled by an explicit formula (Petersen and Pedersen, 2012)

$$\begin{aligned}\frac{\partial}{\partial \theta_i}(\log(\det(\mathbf{Q})) - \log(\det(\mathbf{Q}_{\text{C}}))& \\ &= \text{Tr}(\mathbf{Q}^{-1}\frac{\partial}{\partial \theta_i}\mathbf{Q}) - \text{Tr}(\mathbf{Q}_{\text{C}}^{-1}\frac{\partial}{\partial \theta_i}\mathbf{Q}_{\text{C}}) \\ &= \text{Tr}\left[(\mathbf{Q}^{-1} - \mathbf{Q}_{\text{C}}^{-1})\frac{\partial}{\partial \theta_i}\mathbf{Q}\right].\end{aligned}$$

Then the derivative of the quadratic forms are calculated

$$\begin{aligned}\frac{\partial}{\partial \theta_i}\left(-\frac{1}{2}\vec{y}^{\text{T}}\vec{y} \cdot \tau_{\text{noise}} + \frac{1}{2}\vec{\mu}_{\text{C}}\mathbf{Q}_{\text{C}}\vec{\mu}_{\text{C}}\right)& \\ &= 0 + \frac{\partial}{\partial \theta_i}\left(\frac{1}{2}\vec{y}^{\text{T}}\tau_{\text{noise}}\mathbf{S}\mathbf{Q}_{\text{C}}^{-1}\mathbf{S}^{\text{T}}\tau_{\text{noise}}\vec{y}\right) \\ &= -\frac{1}{2}\vec{y}^{\text{T}}\tau_{\text{noise}}\mathbf{S}\mathbf{Q}_{\text{C}}^{-1}\left(\frac{\partial}{\partial \theta_i}\mathbf{Q}_{\text{C}}\right)\mathbf{Q}_{\text{C}}^{-1}\mathbf{S}^{\text{T}}\tau_{\text{noise}}\vec{y} \\ &= -\frac{1}{2}\vec{\mu}_{\text{C}}^{\text{T}}\left(\frac{\partial}{\partial \theta_i}\mathbf{Q}\right)\vec{\mu}_{\text{C}}.\end{aligned}$$

Combining these gives

$$\begin{aligned}
&\frac{\partial}{\partial\theta_i}\log(\pi(\vec{\theta},\tau_{\text{noise}}|\vec{y})) \\
&\qquad = \frac{\partial}{\partial\theta_i}\log(\pi(\vec{\theta},\tau_{\text{noise}})) + \frac{1}{2}\text{Tr}\left[(\mathbf{Q}^{-1}-\mathbf{Q}_{\text{C}}^{-1})\frac{\partial}{\partial\theta_i}\mathbf{Q}\right] \\
&\qquad\quad - \frac{1}{2}\vec{\mu}_{\text{C}}^{\text{T}}\left(\frac{\partial}{\partial\theta_i}\mathbf{Q}\right)\vec{\mu}_{\text{C}}
\end{aligned}$$

C.2 Derivative with respect to $\log(\tau_{\text{noise}})$

First calculate the derivative of the log-determinants

$$\begin{aligned}
&\frac{\partial}{\partial\log(\tau_{\text{noise}})}\left(N\log(\tau_{\text{noise}}) - \log(\det(\mathbf{Q}_{\text{C}}))\right) \\
&\qquad = N - \text{Tr}\left(\mathbf{Q}_{\text{C}}^{-1}\frac{\partial}{\partial\log(\tau_{\text{noise}})}\mathbf{Q}_{\text{C}}\right) \\
&\qquad = N - \text{Tr}\left(\mathbf{Q}_{\text{C}}^{-1}\mathbf{S}^{\text{T}}\mathbf{S}\cdot\tau_{\text{noise}}\right).
\end{aligned}$$

Then the derivative of the quadratic forms

$$\begin{aligned}
&\frac{\partial\left(-\frac{1}{2}\vec{y}^{\text{T}}\vec{y}\cdot\tau_{\text{noise}} + \frac{1}{2}\vec{\mu}_{\text{C}}\mathbf{Q}_{\text{C}}\vec{\mu}_{\text{C}}\right)}{\partial\log(\tau_{\text{noise}})} \\
&\qquad = -\frac{1}{2}\vec{y}^{\text{T}}\vec{y}\cdot\tau_{\text{noise}} + \frac{\partial}{\partial\log(\tau_n)}\frac{1}{2}\vec{y}^{\text{T}}\tau_{\text{noise}}\mathbf{S}\mathbf{Q}_{\text{C}}^{-1}\mathbf{S}^{\text{T}}\tau_{\text{noise}}\vec{y} \\
&\qquad = -\frac{1}{2}\vec{y}^{\text{T}}\vec{y}\cdot\tau_{\text{noise}} + \vec{y}^{\text{T}}\tau_{\text{noise}}\mathbf{S}\mathbf{Q}_{\text{C}}^{-1}\mathbf{S}\left(\frac{\partial\tau_{\text{noise}}}{\partial\log(\tau_{\text{noise}})}\right)\vec{y} \\
&\qquad\quad - \frac{1}{2}\vec{y}^{\text{T}}\tau_{\text{noise}}\mathbf{S}\mathbf{Q}_{\text{C}}^{-1}\left(\frac{\partial}{\partial\log(\tau_{\text{noise}})}\mathbf{Q}_{\text{C}}\right)\mathbf{Q}_{\text{C}}^{-1}\mathbf{S}^{\text{T}}\tau_{\text{noise}}\vec{y} \\
&\qquad = -\frac{1}{2}\vec{y}^{\text{T}}\vec{y}\cdot\tau_{\text{noise}} + \vec{\mu}_{\text{C}}^{\text{T}}\mathbf{S}^{\text{T}}\vec{y}\cdot\tau_{\text{noise}} - \frac{1}{2}\vec{\mu}_{\text{C}}^{\text{T}}\mathbf{S}^{\text{T}}\mathbf{S}\vec{\mu}_{\text{C}}\cdot\tau_{\text{noise}} \\
&\qquad = -\frac{1}{2}(\vec{y}-\mathbf{A}\vec{\mu}_{\text{C}})^{\text{T}}(\vec{y}-\mathbf{A}\vec{\mu}_{\text{C}})\cdot\tau_{\text{noise}}.
\end{aligned}$$

Together these expressions give

$$\begin{aligned}\frac{\partial \log(\pi(\vec{\theta}, \tau_{\text{noise}}|\vec{y}))}{\partial \log(\tau_{\text{noise}})} \\ = \frac{\partial}{\partial \log(\tau_{\text{noise}})} \log(\pi(\vec{\theta}, \tau_{\text{noise}})) + \frac{N}{2} - \frac{1}{2}\text{Tr}\left[\mathbf{Q}_{\text{C}}^{-1}\mathbf{S}^{\text{T}}\mathbf{S} \cdot \tau_{\text{noise}}\right] \\ - \frac{1}{2}(\vec{y} - \mathbf{A}\vec{\mu}_{\text{C}})^{\text{T}}(\vec{y} - \mathbf{A}\vec{\mu}_{\text{C}}) \cdot \tau_{\text{noise}}\end{aligned}$$

C.3 Implementation

The derivative $\frac{\partial}{\partial \theta_i}\mathbf{Q}_c$ can be calculated quickly since it is a simple functions of θ. The trace of the inverse of a matrix A times the derivative of a matrix B only requires the values of the inverse of A for non-zero elements of B. In the above case the two matrices have the same type of non-zero structure, but it can happen that specific elements in the non-zero structure are zero for one of the matrices. This way of calculating the inverse only at a subset of the locations can be handled as described in Rue and Held (2010).

References

Anderes, E. B. and Stein, M. L. (2008). Estimating deformations of isotropic Gaussian random fields on the plane. *The Annals of Statistics*, pages 719–741.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848.

Bolin, D. (2014). Spatial Matérn fields driven by non-gaussian noise. *Scandinavian Journal of Statistics*, 41(3):557–579.

Bolin, D. and Lindgren, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *The Annals of Applied Statistics*, 5(1):523–550.

Bolin, D. and Wallin, J. (2013). Non-gaussian Mat\'ern fields with an application to precipitation modeling. *arXiv preprint arXiv:1307.6366*.

Bornn, L., Shaddick, G., and Zidek, J. V. (2012). Modeling nonstationary processes through dimension expansion. *Journal of the American Statistical Association*, 107(497):281–289.

Bowman, A. W. and Crujeiras, R. M. (2013). Inference for variograms. *Computational Statistics & Data Analysis*, 66:19–31.

Calder, C. A. (2007). Dynamic factor process convolution models for multivariate space–time data with application to air quality assessment. *Environmental and Ecological Statistics*, 14(3):229–247.

Calder, C. A. (2008). A dynamic process convolution approach to modeling ambient particulate matter concentrations. *Environmetrics*, 19(1):39–48.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226.

Damian, D., Sampson, P. D., and Guttorp, P. (2001). Bayesian estimation of semi-parametric non-stationary spatial covariance structures. *Environmetrics*, 12(2):161–178.

Damian, D., Sampson, P. D., and Guttorp, P. (2003). Variance modeling for non-stationary spatial processes with temporal replications. *Journal of Geophysical Research: Atmospheres*, 108(D24).

Diggle, P., Ribeiro, P., and Justiniano, P. (2007). *Model-based Geostatistics*. Springer New York.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational statistics & data analysis*, 53(8):2873–2884.

Fuentes, M. (2001). A high frequency kriging approach for non-stationary environmental processes. *Environmetrics*, 12(5):469–483.

Fuentes, M. (2002a). Interpolation of nonstationary air pollution processes: a spatial spectral approach. *Statistical Modelling*, 2(4):281–298.

Fuentes, M. (2002b). Spectral methods for nonstationary spatial processes. *Biometrika*, 89(1):197–210.

Fuentes, M. (2005). A formal test for nonstationarity of spatial stochastic processes. *Journal of Multivariate Analysis*, 96(1):30–54.

Fuglstad, G.-A., Lindgren, F., Simpson, D., and Rue, H. (2015). Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy. *Statistica Sinica*, 25:115–133. In press.

Gneiting, T., Raftery, A., Westveld III, A., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133(5):1098–1118.

Haas, T. C. (1990a). Kriging and automated variogram modeling within a moving window. *Atmospheric Environment. Part A. General Topics*, 24(7):1759 – 1769.

Haas, T. C. (1990b). Lognormal and moving window methods of estimating acid deposition. *Journal of the American Statistical Association*, 85(412):950–963.

Haas, T. C. (1995). Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *Journal of the American Statistical Association*, 90(432):1189–1199.

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.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics*, 5:173–190. 10.1023/A:1009666805688.

Higdon, D., Swall, J., and Kern, J. (1999). Non-stationary spatial modeling. *Bayesian statistics*, 6(1):761–768.

Ingebrigtsen, R., Lindgren, F., and Steinsland, I. (2014). Spatial models with explanatory variables in the dependence structure of Gaussian random fields based on stochastic partial differential equations. *Spatial Statistics*, 8:20–38.

Johns, C. J., Nychka, D., Kittel, T. G. F., and Daly, C. (2003). Infilling sparse records of spatial fields. *Journal of the American Statistical Association*, 98(464):796–806.

Jun, M. and Genton, M. (2012). A test for stationarity of spatio-temporal random fields on planar and spherical domains. *Statistica Sincia*, 22:1737–1764.

Kim, H.-M., Mallick, B. K., and Holmes, C. C. (2005). Analyzing nonstationary spatial data using piecewise Gaussian processes. *Journal of the American Statistical Association*, 100(470):653–668.

Kleiber, W. and Nychka, D. (2012). Nonstationary modeling for multivariate spatial processes. *Journal of Multivariate Analysis*, 112(0):76 – 91.

Lindgren, F. and Rue, H. (2008). On the second-order random walk model for irregular locations. *Scandinavian journal of statistics*, 35(4):691–700.

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.

Loader, C. and Switzer, P. (1989). Spatial covariance estimation for monitoring data. Technical Report 133, SIAM Institute for Mathematics and Society.

Neto, J. H. V., Schmidt, A. M., and Guttorp, P. (2014). Accounting for spatially varying directional effects in spatial covariance structures. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(1):103–122.

Nott, D. J. and Dunsmuir, W. T. M. (2002). Estimation of nonstationary spatial covariance structure. *Biometrika*, 89(4):819–829.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2014). A multi-resolution gaussian process model for the analysis of large spatial data sets. *Journal of Computational and Graphical Statistics*. In press.

Nychka, D., Wikle, C., and Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2(4):315–331.

Oehlert, G. W. (1993). Regional trends in sulfate wet deposition. *Journal of the American Statistical Association*, 88(422):pp. 390–399.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506.

Petersen, K. B. and Pedersen, M. S. (2012). The matrix cookbook. Version 20121115.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Rue, H. and Held, L. (2010). Markov random fields. In Gelfand, A., Diggle, P., Fuentes, M., and Guttorp, P., editors, *Handbook of Spatial Statistics*, pages 171–200. CRC/Chapman & Hall, Boca Raton, FL.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Sampson, P. D. (2010). Constructions for nonstationary spatial processes. In Alan E. Gelfand, Peter J. Diggle, M. F. and Guttorp, P., editors, *Handbook of Spatial Statistics*, Handbooks of Modern Statistical Methods, chapter 9, pages 119–130. Chapman & Hall/CRC.

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

Schmidt, A. M., Guttorp, P., and O'Hagan, A. (2011). Considering covariates in the covariance structure of spatial processes. *Environmetrics*, 22(4):487–500.

Schmidt, A. M. and O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):743–758.

Sigrist, F., Künsch, H. R., and Stahel, W. A. (2012). A dynamic nonstationary spatio-temporal model for short term prediction of precipitation. *The Annals of Applied Statistics*, 6(4):1452–1477.

Simpson, D., Illian, J., Lindgren, F., Sørbye, S., and Rue, H. (2011). Going off grid: Computationally efficient inference for log-Gaussian Cox processes. *arXiv preprint arXiv:1111.0641*.

Simpson, D., Lindgren, F., and Rue, H. (2012). In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics*, 23(1):65–74.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41(3/4):pp. 434–449.

Whittle, P. (1963). Stochastic-processes in several dimensions. *Bulletin of the International Statistical Institute*, 40(2):974–994.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.

Zimmerman, D. L. (1993). Another look at anisotropy in geostatistics. *Mathematical Geology*, 25(4):453–470.

Paper III

Interpretable Priors for Hyperparameters for Gaussian Random Fields

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H.

2015. *Technical report*

Main result in submission as part of

Penalising model component complexity: A principled, practical approach to constructing priors

Simpson, D., Martins, T., Riebler, A., Fuglstad, G.-A, Rue, H., and Sørbye, S.

2014

Interpretable Priors for Hyperparameters for Gaussian Random Fields

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

[1]Department of Mathematical Sciences, NTNU, Norway
[2]Department of Statistics, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom
[3]Department of Mathematical Sciences, University of Bath, United Kingdom

Abstract

Gaussian random fields (GRFs) are important building blocks in hierarchical models for spatial data, but there is no practically useful, principled approach for selecting the prior on their hyperparameters. The prior is typically chosen in an *ad-hoc* manner, which lacks theoretical justification, despite the fact that we know that the hyperparameters are not consistently estimable from a single realization and that there is sensitivity to the choice of the prior.

We first use the recent Penalised Complexity prior framework to construct a practically useful, tunable, weakly informative joint prior on the range and the marginal variance for Matérn GRFs with fixed smoothness. We then discuss how to extend this prior to a prior for a non-stationary GRF with covariates in the covariance structure.

Keywords: Bayesian, Gaussian random fields, Spatial models, Priors, Range, Variance, Penalised Complexity, Non-stationary

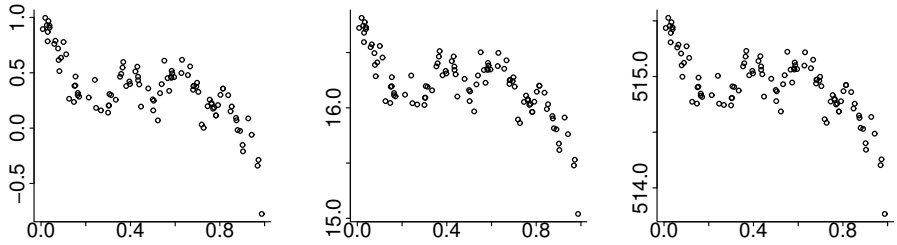
1 Introduction

Gaussian random fields (GRFs) are fundamental building blocks in spatial statistics and non-parametric modelling. They provide a simple and

powerful tool for modelling data with spatial or temporal dependence, but the Gaussian assumption is in many cases too stringent and they are embedded within a hierarchical structure as one of multiple components that controls the behaviour of the observations. In this context, the behaviour of the GRF is usually controlled through a few parameters such as range, marginal variance and smoothness, but, even though GRFs are a standard modelling tool, the choice of prior distribution for the parameters remains a challenge. The prior is difficult to choose: a well-chosen prior will stabilise the inference and improve the predictive performance, whereas a poorly chosen prior can be catastrophic. Due to the infinite-dimensional nature of GRFs, it is difficult to construct a good prior and in most applications the prior is chosen in an *ad-hoc* fashion. In this paper we focus on Matérn GRFs with fixed smoothness, but the methods we develop are more widely applicable.

The lack of practically useful, theoretically founded priors is troubling since there is a ridge in the likelihood along which the value of the likelihood decreases slowly (Warnes and Ripley, 1987), and since the range and the marginal variance for the Matérn family of covariance functions cannot be estimated consistently under in-fill asymptotics (Zhang, 2004). The behaviour of the prior on the ridge will strongly affect the behaviour of the posterior on the ridge and no matter how many points are observed in a bounded observation window, there is a limit to the amount of information that can be learned about these parameters. For example, if a one-dimensional GRF with an exponential covariance function is observed on $[0, 1]$, it is only the ratio of the range and the marginal variance that can be estimated consistently, and not the range and the marginal variance separately (Ying, 1991). When we move along the ridge towards large values of the range and the marginal variance, we change the distribution of the level of the points, but the distribution of the spread around the level changes only slightly. Figure 1 shows how the level moves, but the pattern of the points around the level remains stable for increasing values of the range and the marginal variance.

In some sense, the lack of identifiability is alleviated by the fact that there is a connection between what we can learn from observations and what can affect the predictive distributions, and in this example it is the ratio of the range and the marginal variance that is the important quantity



(a) $\rho = \sigma^2 = 1$ (b) $\rho = \sigma^2 = 1000$ (c) $\rho = \sigma^2 = 1000000$

Figure 1: Simulations with the exponential covariance function $c(d) = \sigma^2 \mathrm{e}^{-d/\rho}$ for different values of $\rho = \sigma^2$ using the same underlying realization of independent standard Gaussian random variables. The patterns of the values are almost the same, but the levels differ.

for the asymptotic properties of the predictions (Stein, 1999). But even though there is a place for intrinsic models in spatial statistics, a practitioner who observes the values in Figure 1a is unlikely to believe that the ranges and marginal variances that can generate Figures 1b and 1c are correct even if the spread is consistent with the observed pattern. In this problem the likelihood by itself is not informative enough to properly control the sizes of the credible intervals, therefore, the practitioner should be provided with a prior that allows control, in an interpretable way, of how far the posterior is allowed to move along the ridge.

Despite this, to our knowledge, the only principled approach to prior selection for GRFs was introduced by Berger et al. (2001), who derived reference priors for a GRF partially observed with no noise. These priors fundamentally depend on the design of the experiment, which makes them inappropriate as "blind" default priors or when data is being analysed in a sequential fashion. This work has been extended by several authors (Paulo, 2005; Kazianka and Pilz, 2012; Kazianka, 2013) – critically Oliveira (2007) allowed for Gaussian observation noise – however, these papers have the same design dependence as the original work. Furthermore, the priors are not applicable as default priors for hierarchical models because the assumption of Gaussian observation noise is insufficient in many situations and

there are currently no extensions of spatial reference priors to other observation processes. In the more restricted case of a GRF with a Gaussian covariance function van der Vaart and van Zanten (2009) showed that the inference asymptotically behaves well with an inverse gamma distribution on range, but they provide no guidance on which hyperparameters should be selected for the prior.

In practice, the range is commonly given a uniform distribution on a bounded interval, a log-uniform distribution on a bounded interval or an inverse gamma distribution with infinite variance and the mean placed at an appropriate location. These priors have little theoretical foundation and are *ad-hoc* choices based on the idea that they will allow reasonable ranges, but Berger et al. (2001) noted that the posterior inference can be sensitive to the choice of cut-off for the uniform prior, and it is necessary with careful sensitivity analysis. The bounded intervals are necessary because improper prior distributions cannot be applied without great care as they tend to lead to improper posterior distributions. From a Bayesian modelling perspective this is an unsatisfactory situation because the prior is supposed to encode the user's uncertainty about the parameters and not be an *ad-hoc* choice made out of convenience without theoretical justification.

We apply the recent Penalised Complexity (PC) prior framework developed by Simpson et al. (2014) to construct a new, principled joint prior for the range and the marginal variance of a Matérn GRF. The PC prior framework ignores the observation process entirely and focuses instead on the geometry of the parameter space induced by the infinite-dimensional GRF. This is more technically demanding than considering only the finite-dimensional observation, like for the reference priors, but we are able to use the resulting prior for any spatial design and any observation process. The second key difference between the reference priors and the PC prior approach is that while the former is "non-informative" in a technical sense, PC priors are weakly informative and, therefore, require specific information from the user. In particular, PC priors need a point in the parameter space, considered a *base model*, and hyperparameters indicating how strongly the user wishes to shrink towards the base model. Simpson et al. (2014) showed that the resulting inference was quite robust against the specification of the hyperparameters.

The reference prior makes the posterior decay slowly along the ridge since predictions are not heavily influenced by the near intrinsicness in the level, but the PC-prior is weakly informative and allows the user to force the posterior to decay quicker along the ridge. When we incorporate a prior belief that the marginal standard deviation is below a specific value, we cannot move much past this value on the ridge without violating the prior belief. In this way it is possible to obtain more realistic parameter estimates and smaller and more meaningful credible intervals than with the reference priors. The reference priors are fully based on the likelihood and have no options for controlling how far the spatial model is allowed to move towards near-intrinsic models with large ranges and large variances even if they do not make sense for the application at hand.

The drawbacks and insufficiencies of noninformative priors for spatial models have already been commented by other authors and the arguments are well summarized by Palacios and Steel (2006) who wrote:

> Thus we need to think carefully about our priors and try to use as much information as we have available in eliciting reasonable prior distributions. In this particular context [Bayesian geostatistical models], we feel that this strategy is preferable to relying on automatic noninformative priors like the reference prior (if such priors are at all available; ...).

The prior for stationary GRFs provides a strong foundation for the development of priors for non-stationary GRFs. The covariance structure of a GRF is only observed indirectly through the values of the process and for locations without observations there is no information about the covariances. Therefore, the estimated covariance structure can be highly model-dependent and it would be useful and important to have an interpretable prior that provides understanding about the *a priori* assumptions that we put into the non-stationary model. We extend the prior for the stationary GRF to a prior for a non-stationary GRF with covariates in the covariance structure. The prior is motivated by the PC prior framework, but has *ad-hoc* components.

We start by deriving the new, joint PC prior for the range and the marginal variance for a Matérn GRF with fixed smoothness parameter in Section 2. Then in Section 3 the frequentist coverage is studied through

a simulation study and compared with the coverage when using the Jeffreys' rule prior and *ad-hoc* uniform and log-uniform priors. In Section 4 we study the behaviour of the joint posterior under the PC-prior and the Jeffreys' rule prior and discuss the difference in behaviour. Then the frequentist properties of spatial logistic regression are studied in Section 5 to demonstrate the applicability of the PC prior for a non-Gaussian observation process. In Section 6 we discuss how to extend the prior for the stationary model to a prior for a non-stationary GRF. The paper ends with discussion and concluding remarks in Section 7.

2 Penalised complexity prior

2.1 Background

The principle idea of the PC-prior framework is to think of a model component as a flexible extension of the *base model*, which is chosen to be the simplest or least flexible state of the model component. For example, a random effect is an extension of a random effect with zero variance, i.e. no random effect. After selecting the base model, one derives a distance measure from the base model to the models described by other parameter values. This distance from the base model describes how much more flexible each model is than the base model and provides a measure of complexity for the model component, and the prior is set directly on the distance from the base model instead of on the parameters of the model. This provides a useful tool for setting priors on parameters for which it is hard to have intuition. For example, correlation parameters close the border values -1, 0 and 1, or the range in spatial models.

To put this idea into practice, it is necessary to decide which measure of complexity to use and which prior to put on the resulting distance. We measure the extra complexity of each model compared to the base model through the Kullback-Leibler divergence (KLD). The KLD of the probability density f from the probability density g is defined by

$$D_{\mathrm{KL}}(f||g) = \int_{\chi} f(\boldsymbol{x}) \log\left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x},$$

and expresses the information lost when g is used to approximate f. The asymmetry of the KLD fits well with the choice of the base model as the

favoured model, and we turn the KLD into a uni-directional distance from the base model g to the model f through $d(f||g) = \sqrt{2\text{KLD}(f||g)}$.

The remaining key point is which distribution to put on the derived distance and Simpson et al. (2014) provide three principles for selecting the prior on the distance: Occam's razor, constant rate penalisation and user-defined scaling. Occam's razor is achieved by constructing a prior that penalises deviations from the base model and favours the base model until the data provides evidence against it. This suggests that the prior density should have its peak at distance 0 and less and less density for higher distances. The constant rate penalisation is achieved by making the prior on the distance, d, satisfy the relationship

$$\frac{\pi(d+\delta)}{\pi(d)} = r^{\delta}, \qquad d, \delta \geq 0,$$

for a constant decay-rate $0 < r < 1$. This means that the relative change in the prior when the distance increases by δ does not depend on the current distance d, and leads to the exponential distribution

$$\pi(d) = \lambda \exp(-\lambda d).$$

After deciding on this distribution we apply the final principle of user-defined scaling to determine the hyperparameter λ. We transform the distance back to an interpretable size $Q(d)$ and include prior information through

$$\text{P}(Q(d) > U) = \alpha \quad \text{or} \quad \text{P}(Q(d) < L) = \alpha,$$

where U or L is an upper or lower limit, respectively, and α is the probability in the upper or lower tail of the prior distribution. By selecting U or L, and α the user combines prior belief with a prior derived from the geometry of the parameter space.

We want to extend the approach outlined above to Gaussian Matérn fields with fixed smoothness. These GRFs have the covariance function

$$C(d) = \sigma^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{\sqrt{8\nu}d}{\rho}\right)^{\nu} K_{\nu}\left(\frac{\sqrt{8\nu}d}{\rho}\right),$$

where ρ is the spatial range, σ^2 is the marginal variance and ν is a fixed smoothness. The first thing we need to decide is what the base model

should be; what type of model do we want to shrink towards? It seems clear that we want to shrink towards zero standard deviation, or no effect, but the GRF is controlled by two parameters and we need to describe how the range behaves as the marginal variance goes to zero. We choose to shrink the range simultaneously towards infinity with the goal of achieving shrinkage towards a constant random field with variance zero.

The GRFs possess a technical difficulty not present for finite dimensional distributions. Imagine that the spatial field is observed at all points in a bounded observation window, for example, $[0, 1]^2$, then the KLD between the distributions specified by two choices of parameters (ρ_0, σ_0^2) and (ρ_1, σ_1^2) is in general infinite. This means that we must be careful in our prior construction and understand which changes corresponds to infinite KLD. To facilitate the construction of the prior, we select a parametrization of the spatial field that more accurately describes what can be and what cannot be estimated from a bounded observation window. We introduce the parameters $\kappa = \sqrt{8\nu}/\rho$ and

$$\tau = \frac{\Gamma(\nu)}{(4\pi)^{d/2}\Gamma(\nu + d/2)\sigma^2\kappa^{2\nu}}.$$

These parameters arise from a slight re-parametrization of the SPDE in Lindgren et al. (2011),

$$(\kappa^2 - \Delta)^{\alpha/2}(\sqrt{\tau}u(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathbb{R}^d, \tag{1}$$

where $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ is the Laplacian and $\mathcal{W}$ is standard Gaussian white noise. If κ and τ are chosen as above, this SPDE specifies a Matérn GRF with range ρ, marginal variance σ^2 and smoothness $\nu = \alpha - d/2$. In this parametrization τ can be consistently estimated from a bounded observation window, whereas κ cannot. If κ is kept fixed and the value of τ changes, the KLD is infinite, but if the value of τ is kept fixed and the value of κ changes, the KLD is finite. This parametrization allows us to use a two-step procedure where we first set a prior on κ based on the distribution of the GRF through the KLD and then set a prior on τ given the value of κ through a consideration of finite-dimensional observations.

2.2 Joint prior on range and marginal variance

We first construct a joint prior for κ and τ through the densities $\pi(\kappa)$ and $\pi(\tau|\kappa)$, and then transform the resulting joint prior to a joint prior on ρ and σ^2. The prior on κ is constructed using the limit of a re-scaled form of the KLD, but the details are technical and the derivation is, therefore, given in Appendix A. The result is a description of how much more complex the models with $\kappa > 0$ are compared to the intrinsic model $\kappa = 0$ for a fixed τ. If we ignore the multiplicative constants, the resulting distance is

$$d(\kappa) = \kappa^{d/2}. \tag{2}$$

The distance in Equation (2) expresses how far the distribution of the GRF is from the intrinsic GRF as a function of κ and solves the highly non-trivial problem of describing how much a Matérn GRF varies as a function of range. Since $\rho = \sqrt{8\nu}/\kappa$, the distance in Equation (2) implies that the range can be made arbitrarily large without making large changes in the distribution if the marginal variance increases in such a way that τ is kept constant. This increase in marginal variance will be controlled by the prior for $\tau|\kappa$, which will disallow unreasonably large marginal variances and thus near-intrinsic models.

An exponential prior on the distance from the base model gives

$$\begin{aligned}\pi(\kappa) &= \lambda_1 \exp(-\lambda_1 d(\kappa)) \left| \frac{\mathrm{d}}{\mathrm{d}\kappa} d(\kappa) \right| \\ &= \frac{\lambda_1 d}{2} \kappa^{d/2-1} \exp(-\lambda_1 \kappa^{d/2}), \quad \kappa > 0,\end{aligned} \tag{3}$$

where λ_1 is determined by controlling the *a priori* probability that the range is below a specific limit,

$$\mathrm{P}\left(\frac{\sqrt{8\nu}}{\kappa} < \rho_0 \right) = \alpha_1,$$

i.e.

$$\lambda_1 = -\log(\alpha_1) \left(\frac{\rho_0}{\sqrt{8\nu}} \right)^{d/2}.$$

The calibration of the prior requires the selection of two values: the lower range, ρ_0, and the probability in the lower tail, α_1.

The prior on τ cannot be derived from the distribution of the process on a bounded observation window since this parameter is completely determined by the values of the process on the observation window. It would be meaningless to put a prior on τ if we observed all values in the observation window, and we must instead choose a situation in which a prior is necessary. We make the assumption that we are interested in observing finite-dimensional quantities from the spatial field.

With κ fixed the joint distribution of a finite number of observations is a multivariate Gaussian distribution of the form

$$\pi(\boldsymbol{u}) \propto \exp\left(-\frac{\tau}{2}\boldsymbol{u}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{u}\right),$$

where Σ is a fixed matrix. In this distribution τ acts as a precision parameter and we can use the prior constructed by (Simpson et al., 2014),

$$\pi(\tau) = \frac{\lambda_2}{2}\tau^{-3/2}\exp(-\lambda_2\tau^{-1/2}), \quad \tau > 0, \tag{4}$$

where λ_2 is determined by controlling the *a priori* probability that the marginal variance exceeds a specific level,

$$\mathrm{P}\left(\frac{C(\nu)}{\tau\kappa^{2\nu}} > \sigma_0^2 \middle| \kappa\right) = \alpha_2, \tag{5}$$

where

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

is the constant needed to make the left-hand side of the inequality equal to the marginal variance of the GRF.

Since the calibration criterion in Equation (5) is conditional on the value of κ, it introduces dependence between κ and τ in the joint prior. We write Equation (5) as

$$\mathrm{P}\left(\tau < \frac{C(\nu)}{\kappa^{2\nu}\sigma_0^2} \middle| \kappa\right) = \alpha_2$$

and find

$$\exp\left(-\lambda_2\left(\frac{C(\nu)}{\kappa^{2\nu}\sigma_0^2}\right)^{-1/2}\right) = \alpha_2,$$
$$\lambda_2 = \frac{\lambda_3}{\kappa^\nu},$$

where λ_3 absorbs the other constants in λ_2. We insert this into Equation (4) and find the conditional distribution

$$\pi(\tau|\kappa) = \frac{\lambda_3\tau^{-3/2}}{\kappa^\nu}\exp(-\lambda_3\kappa^{-\nu}\tau^{-1/2}). \tag{6}$$

This implies that the dependence between κ and τ is affected by the value of the smoothness ν.

The joint prior on κ and τ is found by combining Equation (3) and Equation (6), and is given by

$$\begin{aligned}\pi(\kappa,\tau) &= \pi(\kappa)\pi(\tau|\kappa)\\ &= \frac{\lambda_1\lambda_3 d}{2}\tau^{-3/2}\kappa^{d/2-1-\nu}\exp(-\lambda_1\kappa^{d/2}-\lambda_3\kappa^{-\nu}\tau^{-1/2}).\end{aligned}$$

There is a one-to-one correspondence between κ and τ, and ρ and σ^2,

$$\begin{bmatrix}\rho\\ \sigma^2\end{bmatrix} = \begin{bmatrix}\frac{\sqrt{8\nu}}{\kappa}\\ \frac{C(\nu)}{\kappa^{2\nu}\tau}\end{bmatrix},$$

which can be exploited to transform the joint prior for κ and τ to the joint prior for ρ and σ^2,

$$\pi(\rho,\sigma^2) = \left[\frac{d\lambda_4}{2}\rho^{-1-d/2}\exp\left(-\lambda_4\rho^{-d/2}\right)\right]\left[\frac{\lambda_5}{2}\sigma^{-1}\exp\left(-\lambda_5\sigma\right)\right], \tag{7}$$

where λ_4 and λ_5 are selected according to the *a priori* statements

$$\mathrm{P}(\rho<\rho_0)=\alpha_4 \quad \text{and} \quad \mathrm{P}(\sigma^2>\sigma_0^2)=\alpha_5,$$

which give

$$\lambda_4 = -\rho_0^{d/2}\log(\alpha_4) \quad \text{and} \quad \lambda_5 = -\frac{\log(\alpha_5)}{\sigma_0}.$$

3 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. If a prior is intended as a default prior, it should lead to good frequentist properties such as a frequentist coverage of the equal-tailed $100(1-\alpha)\%$ Bayesian credible intervals that is close to the nominal $100(1-\alpha)\%$. We replicate their simulation study with one key difference: we do not include covariates and measurement noise. The reference priors are not proper distributions and the goal of the series of papers was to derive them for different situations such as a spatial field combined with covariates, and a spatial field combined with covariates and Gaussian measurement noise. However, in this paper we construct a prior for the GRF component itself and we are *not* constructing a prior for the GRF together with covariates or together with covariates and Gaussian measurement noise. This is possible because the PC-prior is a proper distribution and can be applied to a spatial field together with covariates and arbitrary observation processes without worrying about the properness of the posterior.

The study uses an isotropic GRF, u, with an exponential covariance function $c(d) = \exp(-2d/\rho_0)$ observed at the locations shown in Figure 2. The observation locations were randomly selected within the domain $[0,1]^2$ and are distributed in an irregular pattern. The study is performed for two values of the nominal range: a short range, $\rho_0 = 0.1$, and a long range, $\rho_0 = 1$. We generate multiple realizations and for each realization we assume that the field is observed directly and fit the model

$$y_i = u(\boldsymbol{s}_i), \qquad i = 1, 2, \ldots, 25,$$

where u is a GRF with an exponential covariance function with parameters ρ and σ^2. We apply 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). The full expressions for the priors are given in Table 1.

For each choice of prior and hyperparameters we generate 1000 observation vectors $\boldsymbol{y} = (y_1, y_2, \ldots, y_{25})$ and estimate the equal-tailed 95%

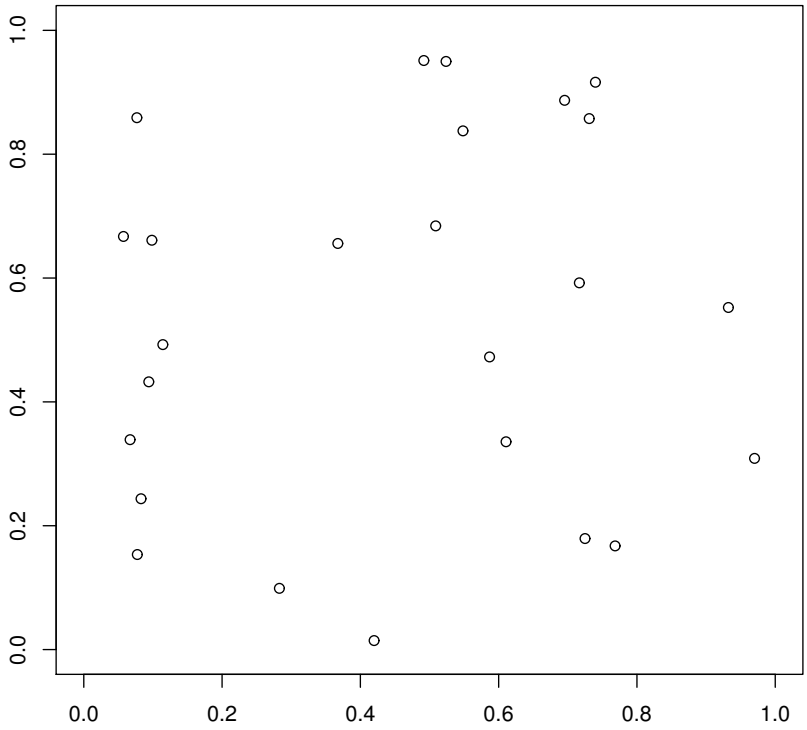


Figure 2: Spatial design for the simulation study of frequentist coverage.

Table 1: The four different priors used in the study of frequentist coverage. 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

credible interval for each observation by running an MCMC-chain. The number of times the true value is contained within the estimated credible interval is divided by 1000 and given as the estimate of the frequentist coverage. We tried MCMC chains of length 25000 with 5000 iterations burn-in and MCMC chains of length 125000 with 25000 burn-in. The results for PriorPC, PriorUn1 and Prior2 were stable, but the chains for PriorJe showed in some cases notoriously high autocorrelation and unstable results and we re-ran with MCMC chains of length 1500000 with 300000 iterations as burn-in.

PriorJe has no hyperparameters, but PriorPC, PriorUn1 and PriorUn2 each has hyperparameters that need to be set before using the prior. 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 a calibration criterion to help choosing the hyperparameters, which helps give an idea about which prior assumptions the chosen hyperparameters are expressing.

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

The results for PriorPC, PriorUn1 and PriorUn2 are shown in Tables 2 and 5, Tables 3 and 6, and Tables 4 and 7, respectively. The results for

Table 2: Frequentist coverage of 95% credible intervals for range and marginal variance when the range $\rho_0 = 0.1$ using PriorPC, where 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	0.768 [0.25]	0.749 [0.24]	0.760 [0.20]	0.693 [0.17]
0.1	0.965 [0.35]	0.976 [0.29]	0.961 [0.27]	0.937 [0.21]
0.4	0.990 [0.45]	0.989 [0.41]	0.993 [0.33]	0.987 [0.25]
1.6	0.717 [0.98]	0.692 [0.82]	0.756 [0.54]	0.807 [0.34]

(b) Marginal variance

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.025	0.941 [1.5]	0.952 [1.4]	0.957 [1.3]	0.918 [0.97]
0.1	0.953 [1.6]	0.966 [1.5]	0.944 [1.4]	0.927 [0.98]
0.4	0.953 [2.0]	0.952 [1.8]	0.960 [1.5]	0.943 [1.1]
1.6	0.904 [3.9]	0.906 [3.2]	0.939 [2.2]	0.972 [1.3]

Table 3: Frequentist coverage of 95% credible intervals for range and marginal variance when the true range $\rho_0 = 0.1$ using PriorUn1, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.901 [0.95]	0.901 [8.6]	0.847 [122]
$5 \cdot 10^{-3}$	0.935 [0.92]	0.918 [7.7]	0.887 [110]
$5 \cdot 10^{-4}$	0.948 [0.93]	0.929 [7.9]	0.893 [110]

(b) Marginal variance

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.952 [3.5]	0.941 [29]	0.895 [460]
$5 \cdot 10^{-3}$	0.945 [3.3]	0.937 [27]	0.907 [410]
$5 \cdot 10^{-4}$	0.953 [3.3]	0.925 [27]	0.921 [412]

Table 4: Frequentist coverage of 95% credible intervals for range and marginal variance when the true range $\rho_0 = 0.1$ using PriorUn2, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.986 [0.47]	0.979 [0.84]	0.988 [1.1]
$5 \cdot 10^{-3}$	0.976 [0.44]	0.950 [0.81]	0.966 [1.0]
$5 \cdot 10^{-4}$	0.932 [0.40]	0.945 [0.70]	0.944 [1.3]

(b) Marginal variance

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.949 [2.0]	0.962 [2.9]	0.965 [3.6]
$5 \cdot 10^{-3}$	0.968 [1.8]	0.960 [2.6]	0.959 [3.2]
$5 \cdot 10^{-4}$	0.948 [1.7]	0.960 [2.4]	0.949 [3.7]

Table 5: Frequentist coverage of 95% credible intervals for range and marginal variance when the true range $\rho_0 = 1$ using PriorPC, where 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	0.950 [12]	0.945 [7.1]	0.906 [3.2]	0.821 [1.4]
0.1	0.977 [15]	0.966 [8.2]	0.962 [3.6]	0.866 [1.5]
0.4	0.965 [26]	0.981 [13]	0.992 [5.1]	0.988 [1.8]
1.6	0.159 [74]	0.349 [31]	0.700 [11]	0.954 [3.3]

(b) Marginal variance

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.025	0.944 [11]	0.956 [6.2]	0.933 [2.8]	0.797 [1.1]
0.1	0.957 [13]	0.966 [7.2]	0.954 [3.1]	0.865 [1.2]
0.4	0.943 [23]	0.957 [11]	0.987 [4.4]	0.972 [1.5]
1.6	0.441 [68]	0.534 [29]	0.797 [9.1]	0.984 [2.5]

Table 6: Frequentist coverage of 95% credible intervals for range and marginal variance when the true range $\rho_0 = 1$ using PriorUn1, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.995 [1.5]	0.831 [18]	0.593 [188]
$5 \cdot 10^{-3}$	0.996 [1.5]	0.818 [18]	0.539 [188]
$5 \cdot 10^{-4}$	0.994 [1.5]	0.844 [18]	0.537 [188]

(b) Marginal variance

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.979 [2.0]	0.857 [20]	0.614 [208]
$5 \cdot 10^{-3}$	0.979 [2.0]	0.821 [20]	0.585 [205]
$5 \cdot 10^{-4}$	0.969 [2.0]	0.828 [20]	0.561 [206]

Table 7: Frequentist coverage of 95% credible intervals for range and marginal variance when the true range $\rho_0 = 1$ using PriorUn2, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.980 [1.5]	0.959 [12]	0.933 [69]
$5 \cdot 10^{-3}$	0.974 [1.5]	0.954 [12]	0.954 [67]
$5 \cdot 10^{-4}$	0.964 [1.5]	0.953 [13]	0.956 [68]

(b) Marginal variance

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.955 [1.8]	0.952 [12]	0.945 [61]
$5 \cdot 10^{-3}$	0.962 [1.8]	0.943 [12]	0.941 [60]
$5 \cdot 10^{-4}$	0.939 [1.8]	0.946 [12]	0.953 [60]

PriorJe was 97.0% coverage with average length of the credible intervals of 0.86 for range and 96.0% coverage and average length of the credible intervals of 2.7 for marginal variance for $\rho_0 = 0.1$, and 95.4% coverage with average length of the credible intervals of 445 for range and 94.4% coverage with average length of the credible intervals of 355 for variance for $\rho_0 = 1$. It is clear from the tables that for PriorPC, PriorUn1 and PriorUn2 the coverage and the length of the credible intervals are dependent on the choice of hyperparameters. This is not surprising since there are few observation and there is a ridge in the likelihood where the behaviour is strongly dependent on the the prior. The length of the credible intervals are, in general, more well-behaved for $\rho_0 = 0.1$ than for $\rho_0 = 1$ because there is more information available about range when the range is short compared to the domain size.

For PriorUn1 the coverage and the length of the credible intervals is strongly dependent on the upper limit in the prior. The prior has the undesirable property of including stronger and stronger prior belief in high ranges when the upper limit is increased. One might argue that the upper limit would never be selected as extreme as in the example, but it verifies the observation of Berger et al. (2001) that the inference is sensitive to the hyperparameters for this prior. For PriorUn2 the coverage is good in both the short range and long range situation, but the lengths of the credible intervals are sensitive to the upper limit of the prior. The new PriorPC exhibit sensitivity in the coverage and the lengths of the credible intervals, but for this prior it is caused by explicitly including information that conflicts with the true value, whereas for PriorUn1 and PriorUn2 it is not immediately clear what information is included through the different choices of hyperparameters.

The coverage of PriorJe is good, but the credible intervals seem excessively long and the prior is more computationally expensive than the other priors. PriorJe is only computationally feasible for low amounts of points since there is a cubic increase in complexity as a function of the number of observations. The average length of the credible intervals for $\rho_0 = 1$ for marginal variance is 355, which imply unreasonably high standard deviations. The high standard deviations do not seem consistent with an observation with values contained between -3 and 3. We study the credible intervals for PriorPC and PriorJe closer for a specific realization in the

next section to gain intuition about why this happen.

With respect to computation time and easy 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. Based on one of the realizations in this simulation study we are unlikely to believe that the spatial field could have a standard deviation greater than 4, and by encoding this information in PriorPC we can limit the upper limits of the credible intervals both for range and marginal variance.

4 Behaviour of the joint posterior

The sensitivity of the length of the credible intervals to the prior and the extreme length of the credible intervals seen for the Jeffreys' rule prior are not entirely surprising due to the ridge in the likelihood, but they are troubling. In the previous section we only looked at properties of the marginal credible intervals, but these do not tell the entire story because there is strong dependence between range and marginal variance in the joint posterior distribution. We study this dependence by studying the posterior distribution for the realization shown in Figure 3. The true range used to simulate the realization is 1. We draw samples from the joint posterior using the PC-prior with parameters $\alpha_\rho = 0.05$, $\rho_0 = 0.1$, $\alpha_\sigma = 0.05$ and $\sigma_0 = 10$, and we draw samples from the joint posterior using the Jeffreys' rule prior.

Figure 4 shows that the upper tails of the posteriors when the Jeffreys' rule prior is used are heavier than the upper tails of the posteriors when the PC-prior is used. The lower endpoints of the credible intervals are similar for both priors, but there is a large difference in the upper limits because the likelihood decays slowly along the ridge and the behaviour of the prior on the ridge is important for the behaviour of the posterior. The marginal posterior distributions do not show the full story about the inference on range and marginal variance because the two parameters are strongly dependent in the posterior distribution. The PC-prior for range has a heavy upper tail for range and the upper tail of the posterior of

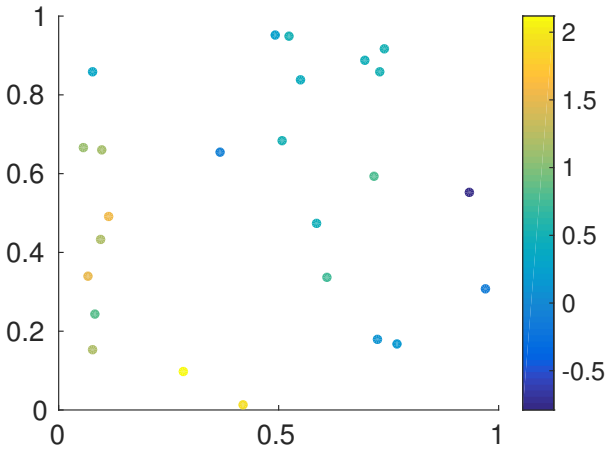
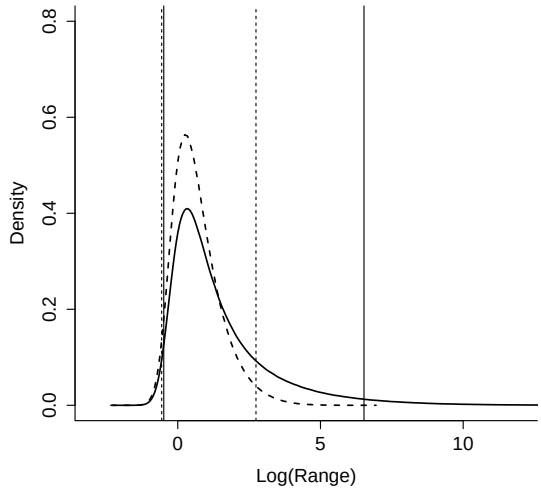


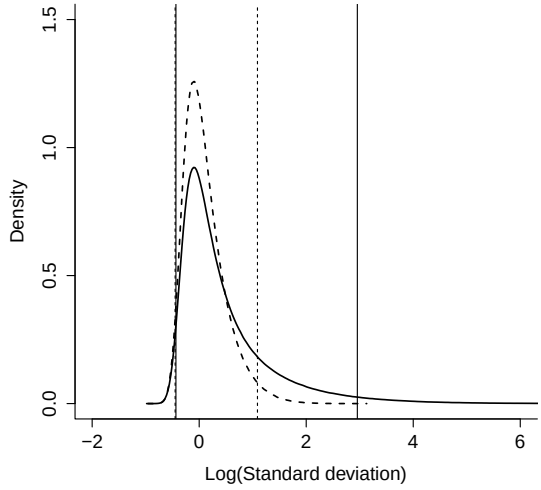
Figure 3: One realization of a GRF with the covariance function $c(d) = \exp(-2d)$ at 25 selected locations.

range is controlled through the prior on marginal variances. The large difference in the marginal posterior for the nominal range in Figure 4a can be explained by the behaviour of the joint posterior.

Figure 5 shows the strong posterior dependence between nominal range and standard deviation in the tail of the distribution. The extreme tail of the Jeffreys' rule prior corresponds to movement far along the ridge in the likelihood. Stein (1999) showed that the ratio of range and marginal variance is the important quantity for asymptotic predictions with the exponential covariance function, which means that long tails are not a major concern for predictions, but for interpretability of range and marginal variance this heavy tail presents a problem. The values of all the observations in Figure 3 lie in the range -1 to 3 and it is unlikely that the true standard deviation should be on the order of 20. After conditioning on data the effect of using a near intrinsic GRF with simultaneously large values for range and marginal variance is almost the same as a GRF with meaningful values for range and marginal variance. Intrinsic models have a place in statistics, but the results show that the Jeffreys' rule prior has the, potentially, undesirable behaviour of favouring intrinsic GRFs with large marginal standard deviations and ranges. The PC prior offers a way to introduce prior belief about the marginal standard deviations, and thus



(a) Posterior for the logarithm of range



(b) Posterior for the logarithm of marginal standard deviation

Figure 4: Marginal posteriors of the logarithm of range and the logarithm of marginal standard deviation. The dashed lines shows the posterior and the credible intervals when the PC-prior is used and the solid line shows the posterior and the credible intervals when the Jeffreys' rule prior is used.

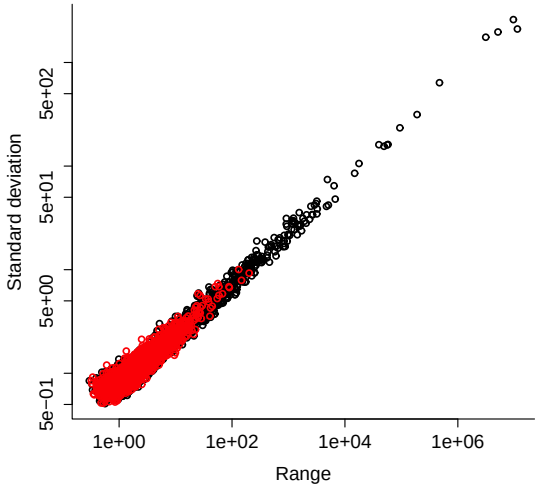


Figure 5: Samples from the joint posterior of range and marginal standard deviation. The red circles are samples using the PC-prior and the black circles are samples using the Jeffreys' rule prior.

a way to avoid the intrinsic GRFs and keep the standard deviation at reasonable (according to prior belief) values.

5 Example: Spatial logistic regression

What makes the PC prior more practically useful than the reference prior, beyond the computational benefits and interpretability, is that the prior is applicable in any hierarchical model and does not have to be re-derived each time a component is removed or added, or the observation process is changed. We consider a simple spatial logistic regression example to demonstrate the applicability of the PC prior beyond direct observations or Gaussian measurement noise.

We select the 25 locations in Figure 2 and generate realizations from the model

$$y_i|p_i \sim \text{Binomial}(20, p_i), \quad i = 1, 2, \ldots, 25,$$

where

$$\text{probit}(p_i) = u(\boldsymbol{s}_i),$$

where u is a GRF with the exponential covariance function with parameters $\rho = 0.1$ and $\sigma = 1$. For each realization the parameters ρ and σ^2 are assumed unknown and must be estimated. The posterior of the parameters is estimated with an MCMC chain and the equal-tailed 95% credible intervals are estimated from the samples of the MCMC-chain after burn-in. We repeat the procedure above 500 times and report the number of times the true value is contained in the credible interval and the average length of the credible interval.

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 8 are similar to the values in Table 2 except that the credible intervals are slightly longer. The longer credible intervals are reasonable since the binomial likelihood gives less information about the spatial field than direct observation of the spatial field. The coverage for 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 3.

6 Priors on non-stationarity

The development of practically useful, interpretable priors for stationary GRFs is important and useful, but the need for such priors is even stronger for non-stationary GRFs. The covariance structure estimated with a non-stationary GRF can be strongly dependent on the *a priori* assumptions on the non-stationarity. It can be difficult to understand the implications of the *a priori* assumptions that we put into non-stationary models because it is difficult to understand how the distribution of a GRF varies as a function of the parameters. The two main challenges are to construct a prior, which accounts for the highly non-trivial geometry of the parameter space, and to calibrate the prior in an interpretable way. Therefore, the PC prior framework is an appealing starting point with properties that fit well for developing such a prior.

There exists different models for non-stationary data and they in-

Table 8: 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
0.025	0.804 [0.29]	0.790 [0.24]	0.774 [0.22]	0.726 [0.19]
0.1	0.974 [0.41]	0.986 [0.37]	0.974 [0.33]	0.956 [0.24]
0.4	0.996 [0.61]	0.982 [0.57]	0.996 [0.43]	0.992 [0.30]
1.6	0.648 [1.4]	0.604 [1.2]	0.722 [0.67]	0.762 [0.44]

(b) Marginal variance

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.025	0.942 [2.0]	0.946 [1.9]	0.948 [1.7]	0.912 [1.2]
0.1	0.920 [2.3]	0.942 [2.0]	0.964 [1.8]	0.922 [1.2]
0.4	0.952 [2.7]	0.962 [2.4]	0.968 [1.9]	0.928 [1.2]
1.6	0.904 [5.3]	0.936 [4.1]	0.966 [2.7]	0.982 [1.5]

corporate non-stationarity in different ways. For example, in the deformation method (Sampson and Guttorp, 1992; Schmidt and O'Hagan, 2003; Damian et al., 2001, 2003) a stationary GRF is made non-stationary through a spatial deformation, in the process convolution method (Haas, 1990b,a; Paciorek and Schervish, 2006) a spatially varying kernel function is convolved with Gaussian white noise, and in the stochastic partial differential equation (SPDE) approach (Bolin and Lindgren, 2011; Fuglstad et al., 2015a,b) a non-stationary GRF is specified through an SPDE with spatially varying coefficients. Each of these types of models has had extensions to covariates in the covariance structure (Schmidt et al., 2011; Neto et al., 2014; Ingebrigtsen et al., 2014a,b).

Ideally, we would derive a prior that could deal with any type of non-stationarity and be applicable for any model for non-stationarity, but, in practice, this is not feasible. For the purpose of this discussion the starting point is the sub-class of the SPDE models (Lindgren et al., 2011) consisting of the model discussed in Ingebrigtsen et al. (2014a). This model uses covariates, and thus needs fewer parameters and is less computationally expensive than a model with a more flexible covariance structure. The model is an extension of the stationary SPDE in Equation (1) with a slightly different parametrization and coefficients that vary spatially,

$$[\kappa(\boldsymbol{s})^2 - \Delta](\tau(\boldsymbol{s})u(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \qquad \mathcal{D} \subset \mathbb{R}^d, \tag{8}$$

with Neumann boundary conditions. We have fixed $\alpha = 2$ to get a practically feasible model, but with a spatially varying range it is unlikely to pose a large practical limitation to fix the smoothness $\nu = 1$.

We make the assumption that the priors on the correlation structure and the marginal variances can be set independently in an analogous way to the stationary GRF. This means we must solve two challenges: covariates must be included separately in the correlation structure and in the marginal variances, and practically useful, interpretable priors must be developed for the covariates in the correlation structure and for the covariates in the marginal variances. Thus setting priors on non-stationarity is not only a question about which prior to set after the parametrization is decided, but a question of how to parametrize the non-stationarity *and* how to set priors on the parameters in the parametrization.

6.1 Parametrizing the non-stationarity

Ingebrigtsen et al. (2014a) expands $\log(\kappa(\cdot))$ and $\log(\tau(\cdot))$ in Equation (1) into low-dimensional bases, but experience numerical problems and prior sensitivity to the priors for the weights in the basis expansions. Ingebrigtsen et al. (2014b) attempt to solve this by setting the hyperparameters of the priors based on the properties of the spatially varying local ranges and marginal variances. The procedure improves the calibration step of the prior specification compared to Ingebrigtsen et al. (2014a), but does not solve the inherent problem that $\kappa(\cdot)$ affects both the correlation structure and the marginal variances of the spatial field. We aim to improve their procedure by first improving the parametrization of the non-stationarity, and then setting and calibrating the prior using the improved parametrization.

The model used by Ingebrigtsen et al. (2014a) introduces spatial variation in the covariance structure by varying the coefficients of the SPDE, but there exists another way to introduce non-stationarity. Instead of varying the coefficients of the SPDE, one can vary the geometry of the space in a similar way as the deformation method. If E is the Euclidean space $\mathbb{R}^2$, the simple SPDE

$$(1 - \Delta_E)u(\boldsymbol{s}) = \sqrt{4\pi}\mathcal{W}_E(\boldsymbol{s}), \quad \boldsymbol{s} \in E, \tag{9}$$

generates a stationary Matérn GRF with range $\rho = \sqrt{8}$, marginal variance $\sigma^2 = 1$ and smoothness $\nu = 1$. Instead of introducing spatially varying coefficients, we introduce spatially varying distances in the space on which the SPDE is defined. We take a two-dimensional manifold $E = \mathbb{R}^2$ and give 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 means that distances are locally scaled by a factor $R(\boldsymbol{s})^{-1}$, or more specifically,

$$\mathrm{d}\sigma^2 = \begin{bmatrix}\mathrm{d}s_1 & \mathrm{d}s_2\end{bmatrix} \mathbf{g}(\boldsymbol{s}) \begin{bmatrix}\mathrm{d}s_1 \\ \mathrm{d}s_2\end{bmatrix} = R(\boldsymbol{s})^{-2}(\mathrm{d}s_1^2 + \mathrm{d}s_2^2), \tag{10}$$

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

The line element in two-dimensional Euclidean space $\sqrt{\mathrm{d}s_1^2 + \mathrm{d}s_2^2}$ is everywhere scaled according to the function $R(\cdot)$, and Equation (10) describes the non-stationary through a spatially varying geometry, which

results in a curved two-dimensional manifold that must be embedded dimension higher than 2 to exist in Euclidean space. The SPDE is not stationary on this space and does not lead to constant marginal variance because the curvature of the space is non-constant unless $R(\cdot)$ is a constant function, but there will be less interaction between $R(\cdot)$ and the marginal variance than $\kappa(\cdot)$ and the marginal variance. And for a slowly varying $R(\cdot)$ the variation in marginal variances is small.

The above construction gives geometric intuition about what type of non-stationarity the equation can generate, but it is not directly useful for implementation. 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 elements of the spaces differ, $\mathrm{d}V_E = \sqrt{\det(g)} \mathrm{d}V_{\mathrm{R}^2}$. We use the SPDE

$$(R(s)^{-2} - \Delta_{\mathbb{R}^2}) u(\boldsymbol{s}) = \sqrt{4\pi} R(s)^{-1} \mathcal{W}_{\mathbb{R}^2}, \quad \boldsymbol{s} \in \mathbb{R}^2, \tag{11}$$

in Euclidean space, but can interpret the SPDE through the implied metric tensor. The SPDE is similar to setting $\kappa(\cdot) = R(\cdot)^{-1}$ in Equation (8), but has an extra factor on the right-hand side of the equation to reduce the variability of the marginal variances.

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 according to the local range function

$$R(s_1, s_2) = \begin{cases} 1 & 0 \leq s_1 < 3, 0 \leq s_2 \leq \pi, \\ (s_1 - 2) & 3 \leq s_1 < 6, 0 \leq s_2 \leq \pi, \\ 4 & 6 \leq s_1 \leq 9, 0 \leq s_2 \leq \pi, \end{cases} \tag{12}$$

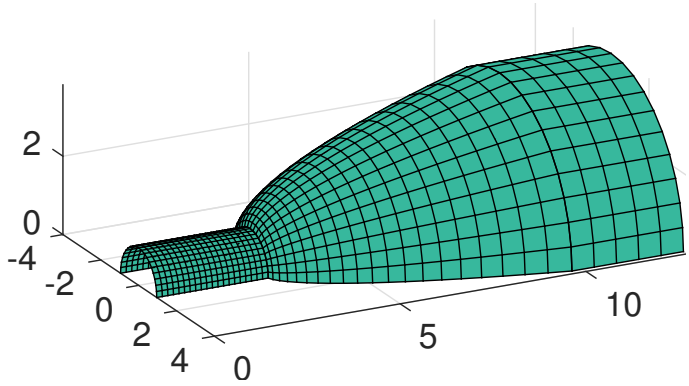


Figure 6: Half cylinder deformed according to the spatially varying metric tensor. The lines formed a regular grid on the half cylinder before deformation.

the space cannot be embedded in $\mathbb{R}^2$. With this metric tensor, the space is no longer flat, but it can be embedded in $\mathbb{R}^3$ as the deformed cylinder shown in Figure 6. Thus, solving Equation (11) with the spatially varying coefficient is the same as solving Equation (9) 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.

However, the SPDE does not completely describe a deformation since solving the "stationary" SPDE on a curved space leads to changes in the marginal variances, but if $R(\cdot)$ does not vary too much, the marginal variances are close to 1. We can, therefore, introduce a separate function $S(\cdot)$ that controls the marginal variances of the process and limit the SPDE to a region of interest, $\mathcal{D}$, with Neumann boundary conditions,

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

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 is different than the SPDE in Equation (8) beyond a reparametrization, and allows for greater separation of the parameters that

affect correlation structure and the parameters that affect marginal variances than the SPDE in Equation (8). This demonstrates that even though two SPDEs are similar and able to capture similar types of behaviour, one can be more useful for setting priors. The SPDE derived based on the metric tensor allows for separate priors for correlation structure and marginal variances through expansions of $\log(R(\cdot))$ and $\log(S(\cdot))$ into bases.

6.2 Setting priors on the non-stationarity

A stationary GRF described through a range ρ and a marginal variance σ^2 will constitute the base model when we work with non-stationarity. We want to shrink the non-stationary GRF towards the stationary GRF that has the PC prior developed in Section 2 for the range and the marginal variance. Denote the parameters that describe the departure from the base model by $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = \mathbf{0}$ corresponds to the stationary GRF. Following the idea of the PC prior framework, we want to give the "distance" from stationarity a prior conditional on the current stationary model $\pi(\boldsymbol{\theta}|\rho, \sigma^2)$. The construction will be based on the ideas of the PC prior framework, but will not be based on a distance calculated from a formal measure of complexity, and the prior will be an *ad-hoc* prior that is motivated by theoretical principles.

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

$$\begin{aligned} \log(R(\boldsymbol{s})) &= \log\left(\frac{\rho}{\sqrt{8}}\right) + \sum_{i=1}^{n_1} \theta_{1,i} f_{1,i}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D}, \\ \log(S(\boldsymbol{s})) &= \log(\sigma^2) + \sum_{i=1}^{n_2} \theta_{2,i} f_{2,i}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D}, \end{aligned} \tag{13}$$

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 variances 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 local ranges and $\boldsymbol{\theta}_2$ controls the marginal variances.

A simple way to account for different scales and dependencies among the basis functions is to give the non-stationary effect in the correlation

structure and the non-stationary effect in the marginal variances independent g-priors (Zellner, 1986) with $g = 1$,

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

where S_1 is the Gramian,

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

and S_2 is the Gramian,

$$S_{2,i,j} = \langle f_{2,i}, f_{2,j} \rangle_{\mathcal{D}}, \quad \text{for } i, j = 1, 2, \ldots, n_2.$$

In this set-up the Gramians account for the structures of the basis functions and the strengths of the effects are reduced to two precisions parameters τ_1 and τ_2. We choose to give the precision parameters the PC prior for precision parameters for Gaussian distributions developed by Simpson et al. (2014), which is designed to shrink towards the base model of zero effect. Because of our *a priori* ansatz of *a priori* independence between the correlation structure and the marginal variances, we set independent priors

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

In this way we have implicitly described the effect in the correlation structure and the effect in the marginal variances through an *ad-hoc* distance and shrunk the effects towards the base model, which is stationarity.

We calibrate the priors based on the *a priori* relative variations they allow for the local range and for the marginal variance through the *a priori* statements,

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

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

These statements are only based on the relative differences in the local range and the marginal variance from a stationary model, and we can see from Equation (13) that the relative differences do not depend on the

parameters of the stationary model. This means that the prior on the non-stationary GRF separates as

$$\begin{aligned}\pi(\rho, \sigma^2, \boldsymbol{\theta}) &= \pi(\rho)\pi(\sigma^2)\pi(\boldsymbol{\theta}|\rho, \sigma^2) \\ &= \pi(\rho)\pi(\sigma^2)\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ &= \pi(\rho)\pi(\sigma^2)\pi(\boldsymbol{\theta}_1)\pi(\boldsymbol{\theta}_2),\end{aligned}$$

where the first equality uses the conditional prior constructed, the second equality uses that the calibration does not introduce dependence with the parameters for the stationary part of the GRF, and the last equality uses that the priors on the spatially varying part of the local range and the marginal variance are independent.

The above conditions control the probability that the relative differences from the stationary model in the local range and the marginal variance exceed pre-specified levels, and allow the user to control the priors based on beliefs about the variability expected in the local range and the marginal variance. The calibration is slightly different than the one used by Ingebrigtsen et al. (2014b). The approach derived is fundamentally *ad-hoc*, but makes several theoretical improvements over the approach in Ingebrigtsen et al. (2014b) due to the new form of the SPDE that reduces the interaction between the correlation structure and the marginal variances in the prior specification, and due to the use of priors on the effects that follow the PC prior framework principle of shrinking towards the base model.

7 Discussion

In this paper we have presented an answer to an important, open question in Bayesian spatial statistics that previously had no satisfactory answer: which prior should we put on the range and the marginal variance for a Matérn GRF? The range and the marginal variance have seemingly clear interpretations and ideally one would hope that it were possible to infer them from a single observation of the spatial field, but in reality there are no consistent estimators of the range and the marginal variance under in-fill asymptotics and the posterior distributions do not contract even for a complete observation of the process in a bounded observation window.

There is a ridge in the likelihood where the posterior distribution of the parameters always will be affected by the prior on the parameters.

For Matérn GRFs objectivity, through noninformative priors such as the reference priors, is not necessarily the correct answer because it can lead to posterior inference that is not sensible. For in-sample predictions, the near-intrinsic models do not negatively affect predictions and it is possible to use the reference priors if they are applicable for the model. However, if the purpose is to infer the range and the marginal variance, to do out-of-sample predictions, or to generate new scenarios from the model using the posterior distribution of the parameters as the prior distribution, the objective approach can lead to meaningless answers that are not compatible with subjective beliefs about the model. The three GRFs in Figure 1 are similar if they are used to predict the unobserved values within the interval $[0, 1]$, but they are highly different if we want to understand the process that generated the data or use this process to generate new data using the parameter posterior.

In practice, we are likely to have subjective knowledge making high marginal variances unreasonable even when the value of the range is high. With the PC-prior developed in this paper this knowledge can be combined with the geometry of the parameter space through two statements of prior belief about the spatial field. In this way it is possible to not just encode the information about geometry contained in the likelihood, but also prior belief about the range and the marginal variance. For example, it is possible to encode prior belief that the marginal standard deviations are unlikely to exceed a specific upper limit. This information disallows the near-intrinsic models far along the ridge of the likelihood and lead to shorter and more meaningful credible intervals than an objective prior such as the reference prior.

The PC prior is weakly informative and like all subjective priors there is a danger that putting prior mass in the wrong place can negatively affect the inference. The calibration of the PC-prior will by design affect the posterior distribution, but since the prior is not just an *ad-hoc* choice, the hyperparameters of the prior have a clearly defined meaning through information that could potentially be elicited from experts. The study of frequentist properties showed that there were negative consequences of being one order wrong in the prior specification, but the examples only had

25 observations and in realistical settings there will be more information available than this, and with more observations, the sensitivity to the prior specification is likely to be less severe. Further, it is when the true range is long compared to the domain size that the likelihood is insufficient for getting physically meaningful estimates and that it is most important to limit the behaviour of the posterior through an interpretable prior.

The greatest benefits of the PC-prior over the reference priors are that there is no dependence on the sampling design, it works with any observation process and can be used in hierarchical models, it is easy to implement, and it is computationally cheap. The benefits over the *ad-hoc* priors is that it has theoretical justification, and that the hyperparameters are interpretable and connected to prior belief about the scale that the parameters are on. This makes the PC prior useful in practice as opposed to the impractical reference priors, while at the same time having the theoretically justification that is lacked by the practical, but *ad-hoc* priors.

The PC prior is extended to a prior for a non-stationary GRF based on an SPDE model using ideas from the PC prior framework, but the extension has *ad-hoc* elements and is specialized to a GRF that can de-couple the parameters controlling the correlation structure and the parameters controlling the marginal variances. It would be desirable to derive a distance from stationarity to non-stationarity by using the KLD in a similar way as for the stationary GRF and not be restricted to a specific non-stationary GRF, but this is difficult because certain properties of the covariance structure of GRFs are identifiable under in-fill asymptotics while others are not.

For SPDE models, a seemingly desirable way to be independent of the parametrization would be to restrict oneself to the approximation of the model used for computations. In computations, one uses a finite-dimensional approximation of the GRF derived through a finite element approximation on a triangulation of the domain. The multivariate Gaussian distribution of the values at the nodes in the triangulation completely describes the distribution of the approximation of the spatial field and we can compute the distances between non-stationary spatial fields by computing the KLDs between multivariate Gaussian distributions.

However, if we do this, a change in $\kappa(\cdot)$ can be handled, but a change

in $\tau(\cdot)$ makes the KLD diverge to infinity as the mesh is refined. Thus, we can use the KLD to construct a prior for the parameters in $\kappa(\cdot)$, but there is still work left in understanding how the KLD behaves as a function of parameters in $\tau(\cdot)$ when the mesh is refined. If one can understand more about the identifiability of $\tau(\cdot)$ as the mesh is refined, one can use this knowledge to re-scale the KLD in a meaningful way, and separate out the constants and the asymptotic behaviour as a function of $\tau(\cdot)$, but this remains an unsolved challenge.

For stationary GRFs, the paper makes significant progress by finding sensible priors for Matérn GRFs through a practically useful, weakly informative joint prior on range and marginal variance. The important remaining question is shared with the other priors derived with the PC prior framework, namely, how easy is it for a user to set the hyperparameters? The hyperparameters have a clear connection to understandable quantities, but the users are still required to gain intuition about setting priors based on the probability of exceeding or being below a chosen value. This is a question we expect to gain experience with when the new prior is implemented within the INLA package (Rue et al., 2009). For non-stationary GRFs the paper makes progress by providing a motivated, but *ad-hoc* construction of a prior. However, a construction fully based on the PC prior framework remains future work.

A The theoretical details for the derivation of the prior for κ

If we consider the distribution of a Matérn GRF, parametrized through $\kappa = \sqrt{8\nu}/\rho$ and

$$\tau = \frac{\Gamma(\nu)}{(4\pi)^{d/2}\Gamma(\nu + d/2)\sigma^2\kappa^{2\nu}},$$

where ρ is the range, σ^2 is the marginal variance, and ν is the smoothness, on a bounded observation window, the KLD between the distributions for two choices of parameters $\kappa, \kappa_0 > 0$ is always finite and it is possible to use the KLD to describe how different the distributions are. The distributions are not absolutely continuous with respect to a Lebesgue measure and we need to describe the KLD in terms of measures. The KLD of the probability measure Q from the probability measure P is defined by

$$D_{\mathrm{KL}}(P||Q) = \int_{\chi} \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P, \tag{14}$$

where $\mathrm{d}P/\mathrm{d}Q$ is the Radon-Nikodym derivative of P with respect to Q, and expresses the information lost when Q is used to approximate P.

We base the constructions in this section on spectral densities and not directly on covariance functions. Fix τ and ν and consider the Matérn GRF u_κ for different values of κ. Lindgren et al. (2011) showed that this GRF can be expressed as a solution of the stochastic partial differential equation (SPDE)

$$(\kappa^2 - \Delta)^{\alpha/2}(\sqrt{\tau}u_\kappa(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathbb{R}^d, \tag{15}$$

where $\alpha = \nu + d/2$ and $\mathcal{W}$ is standard Gaussian white noise. The SPDE can be used to show that the spectral density of u_κ is given by

$$f_\kappa(\boldsymbol{w}) = \left(\frac{1}{2\pi}\right)^d \frac{1}{\tau(\kappa^2 + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w})^\alpha}. \tag{16}$$

The continuous spectrum in Equation (16) is difficult to use directly and we make an intermediate step through a periodic approximation of the GRF. Restrict SPDE (15) to the domain $\mathcal{D} = [-L/2, L/2]^d$ and apply

periodic boundary conditions. This leads to an approximation $\tilde{u}_\kappa$ of u_κ that can be written as

$$\tilde{u}_\kappa(\boldsymbol{s}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} z_{\boldsymbol{k}} \mathrm{e}^{i\langle 2\pi\boldsymbol{k}/L, \boldsymbol{s}\rangle},$$

where $\{z_{\boldsymbol{k}}\}$ are independent Gaussian random variables with variances given by

$$\begin{aligned}\lambda_{\boldsymbol{k}}(\kappa) &= \frac{1}{\tau(\kappa^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha} \frac{\mathrm{Var}[\langle \mathcal{W}, \mathrm{e}^{i\langle 2\pi\boldsymbol{k}/L,\boldsymbol{s}\rangle}\rangle_{\mathcal{D}}]}{\langle \mathrm{e}^{i\langle 2\pi\boldsymbol{k}/L,\boldsymbol{s}\rangle}, \mathrm{e}^{i\langle 2\pi\boldsymbol{k}/L,\boldsymbol{s}\rangle}\rangle_{\mathcal{D}}^2} \\ &= \frac{1}{\tau(\kappa^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha} \frac{L^d}{L^{2d}} \\ &= \frac{1}{L^d} \frac{1}{\tau(\kappa^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha}. \end{aligned} \tag{17}$$

Using this approximation we calculate the KLD between $\tilde{u}_\kappa$ and $\tilde{u}_{\kappa_0}$ (based on Bogachev (1998, Thm. 6.4.6)),

$$\begin{aligned} 2\mathrm{KLD}(\kappa||\kappa_0) &= \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \left[\frac{\lambda_{\boldsymbol{k}}(\kappa_0)}{\lambda_{\boldsymbol{k}}(\kappa)} - 1 - \log\frac{\lambda_{\boldsymbol{k}}(\kappa_0)}{\lambda_{\boldsymbol{k}}(\kappa)}\right] \\ &= \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \left[\frac{(\kappa_0^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha}{(\kappa^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha} - 1 - \log\frac{(\kappa_0^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha}{(\kappa^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha}\right], \end{aligned} \tag{18}$$

which is a simple expression involving only the spectral densities of the processes. If we add scaling with step-size, it becomes a Riemann sum, and we can write

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

where $E(L, \kappa_0)$ is the error in the Riemann sum.

Since we want the base model $\kappa_0 = 0$, which corresponds to infinite range, we need to be careful about how the error $E(L, \kappa_0)$ behaves as $\kappa_0 \to 0$. If L is fixed, the zero frequency gives an infinite term in the summand. Thus the rate at which L tends to infinity must be related to the rate at which κ_0 tends to zero. If the summand for $\boldsymbol{k} = 0$ tends to zero, the Riemann sum converges and $E(L, \kappa_0) \to 0$. The zero-frequency term

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

converges to zero if $L = o(\kappa_0^{-1})$. We apply this relationship between L and κ_0 and introduce the scaled KLD

$$\begin{aligned}
\tilde{\mathrm{KLD}}(\kappa||0) &= \lim_{\kappa_0 \to 0} \left(\frac{2\pi}{L}\right)^d \mathrm{KLD}(\kappa||\kappa_0) \\
&= \frac{1}{2} \int_{\mathbb{R}^d} \left[\frac{(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w})^\alpha}{(\kappa^2 + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w})^\alpha} - 1 - \log \frac{(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w})^\alpha}{(\kappa^2 + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w})^\alpha}\right] \mathrm{d}\boldsymbol{w}.
\end{aligned}$$

We perform the change variables $\boldsymbol{w} = \kappa \boldsymbol{y}$ and find

$$\begin{aligned}
\tilde{\mathrm{KLD}}(\kappa||0) &= \frac{1}{2} \int_{\mathbb{R}^d} \left[\frac{(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y})^\alpha}{(1 + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y})^\alpha} - 1 - \log \frac{(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y})^\alpha}{(1 + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y})^\alpha}\right] \kappa^d \mathrm{d}\boldsymbol{y} \\
&= \kappa^d \tilde{\mathrm{KLD}}(1||0) \\
&\propto \kappa^d, \qquad (19)
\end{aligned}$$

if $\tilde{\mathrm{KLD}}(1||0)$ exists.

However, $\tilde{\mathrm{KLD}}(1||0)$ does not exist for all dimensions d. Perform a change of coordinates to n-dimensional spherical coordinates to find

$$\tilde{\mathrm{KLD}}(1||0) = C_d \int_0^\infty \left[\left(\frac{r^2}{1+r^2}\right)^\alpha - 1 - \log\left(\frac{r^2}{1+r^2}\right)^\alpha\right] r^{d-1} \mathrm{d}r, \qquad (20)$$

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

$$\tilde{\mathrm{KLD}}(1||0) \leq -C_1 \alpha \int_0^\infty \log \frac{r^2}{1+r^2} \mathrm{d}r = \pi \alpha C_1,$$

and we can conclude that the behaviour around 0 is not a problem for any $d \geq 1$. The behaviour for large r can be studied through an expansion in $(1 + r^2)^{-1}$. The integrand in Equation (20) 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 we can find an $0 < r_0 < \infty$ such that

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

where $C \geq 0$ is a constant. For $d \leq 3$ both terms on the right hand side are finite and based on this and the boundedness for $d = 1$, we can conclude that $\tilde{\mathrm{KLD}}(1||0)$ is finite for $d \leq 3$.

B Calculation of the Kullback-Leibler divergence for a one-dimensional GRF with exponential covariance function

B.1 Goal

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

$$c(d) = \frac{1}{2\kappa} \mathrm{e}^{-\kappa d}, \tag{21}$$

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 $[0, L]$

B.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

$$\mathbf{Q}_\kappa = \frac{1}{\sigma_\kappa^2}\begin{bmatrix} 1 & -\mathrm{e}^{-\kappa\Delta t} & & & & \\ -\mathrm{e}^{-\kappa\Delta t} & 1+\mathrm{e}^{-2\kappa\Delta t} & -\mathrm{e}^{-\kappa\Delta t} & & & \\ & \ddots & \ddots & \ddots & & \\ & & -\mathrm{e}^{-\kappa\Delta t} & 1+\mathrm{e}^{-2\kappa\Delta t} & -\mathrm{e}^{-\kappa\Delta t} & \\ & & & -\mathrm{e}^{-\kappa\Delta t} & 1 \end{bmatrix}. \tag{22}$$

B.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

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

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.

B.3.1 Step 1

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

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

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

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

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

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

B.3.2 Step 2

The determinant of the matrix in Equation (22) 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,

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

Thus the limit is

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

B.4 Full KLD

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

$$\begin{aligned}\mathrm{KLD}(\kappa||\kappa_0) &= \frac{1}{2}\left[\frac{\kappa_0}{\kappa} - 1 + L\left(\frac{\kappa_0^2}{2\kappa} - \frac{\kappa}{2}\right) - \log\left(\frac{\kappa_0}{\kappa}\right) - L(\kappa_0 - \kappa)\right] \\ &= \frac{1}{2}\left[\frac{\kappa_0}{\kappa} - 1 - \log\left(\frac{\kappa_0}{\kappa}\right) + L\left(\frac{\kappa_0^2}{2\kappa} - \kappa_0 + \frac{\kappa}{2}\right)\right]. \qquad (23)\end{aligned}$$

B.5 Comparison with the integral expression

The integral in Appendix A gives the expression

$$\frac{1}{2}\int_{-\infty}^{\infty}\left(\left(\frac{\kappa_0^2 + w^2}{\kappa^2 + w^2}\right) - 1 - \log\left(\frac{\kappa_0^2 + w^2}{\kappa + w^2}\right)\right)\mathrm{d}w = \pi\left(\frac{\kappa_0^2}{2\kappa} - \kappa_0 + \frac{\kappa}{2}\right).$$

If we divide by 2π, this is the same expression as the one that is multiplied with L in Equation (23). This is what we would expect because the integral is derived under the assumption that $L \gg 1/\kappa_0$ and "absorbs" the constant $2\pi/L$.

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. American Mathematical Soc.

Bolin, D. and Lindgren, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *The Annals of Applied Statistics*, 5(1):523–550.

Damian, D., Sampson, P. D., and Guttorp, P. (2001). Bayesian estimation of semi-parametric non-stationary spatial covariance structures. *Environmetrics*, 12(2):161–178.

Damian, D., Sampson, P. D., and Guttorp, P. (2003). Variance modeling for non-stationary spatial processes with temporal replications. *Journal of Geophysical Research: Atmospheres*, 108(D24).

Fuglstad, G.-A., Lindgren, F., Simpson, D., and Rue, H. (2015a). Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy. *Statistica Sinica*, 25:115–133.

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2015b). Does non-stationary spatial data always require non-stationary random fields? *arXiv preprint arXiv:1409.0743*.

Haas, T. C. (1990a). Kriging and automated variogram modeling within a moving window. *Atmospheric Environment. Part A. General Topics*, 24(7):1759 – 1769.

Haas, T. C. (1990b). Lognormal and moving window methods of estimating acid deposition. *Journal of the American Statistical Association*, 85(412):950–963.

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

Ingebrigtsen, R., Lindgren, F., Steinsland, I., and Martino, S. (2014b). Estimation of a non-stationary model for annual precipitation in southern norway using replicates of the spatial field. *arXiv preprint arXiv:1412.2798*.

Kazianka, H. (2013). Objective Bayesian analysis of geometrically anisotropic spatial data. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(4):514–537.

Kazianka, H. and Pilz, J. (2012). Objective Bayesian analysis of spatial data with uncertain nugget and range parameters. *Canadian Journal of Statistics*, 40(2):304–327.

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.

Neto, J. H. V., Schmidt, A. M., and Guttorp, P. (2014). Accounting for spatially varying directional effects in spatial covariance structures. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(1):103–122.

Oliveira, V. d. (2007). Objective Bayesian analysis of spatial data with measurement error. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 35(2):pp. 283–301.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506.

Palacios, M. B. and Steel, M. F. J. (2006). Non-gaussian Bayesian geostatistical modeling. *Journal of the American Statistical Association*, 101(474):604–618.

Paulo, R. (2005). Default priors for Gaussian processes. *The Annals of Statistics*, 33(2):556–582.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

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

Schmidt, A. M., Guttorp, P., and O'Hagan, A. (2011). Considering covariates in the covariance structure of spatial processes. *Environmetrics*, 22(4):487–500.

Schmidt, A. M. and O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):743–758.

Simpson, D. P., Martins, T. G., Riebler, A., Rue, H., Fuglstad, G.-A., and Sørbye, S. H. (2014). Penalising model component complexity: A principled, practical approach to constructing priors. *arXiv preprint arXiv:1403.4630*.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer.

van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675.

Warnes, J. and Ripley, B. (1987). Problems with likelihood estimation of covariance functions of spatial gaussian processes. *Biometrika*, 74(3):640–642.

Ying, Z. (1991). Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *Journal of Multivariate Analysis*, 36(2):280 – 296.

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.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.