Master's thesis

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

Mathias Leander Isaksen

Comparing Global and Local Specification of Spatially Varying Anisotropy

Master's thesis in Applied Physics and Mathematics
Supervisor: Geir-Arne Fuglstad
June 2020

NTNU
Norwegian University of
Science and Technology

Mathias Leander Isaksen

Comparing Global and Local Specification of Spatially Varying Anisotropy

Master's thesis in Applied Physics and Mathematics
Supervisor: Geir-Arne Fuglstad
June 2020

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

NTNU
Norwegian University of
Science and Technology

Preface

This report constitutes my master's thesis at the Norwegian University of Science and Technology (NTNU), and concludes my five years as a student in the Applied Physics and Mathematics programme. It was written during the spring of 2020, and builds upon the work carried out in my specialization project, where a method for non-stationary modeling was investigated. Here, we again consider the same method, and compare it with another non-stationary method.

I would like to thank my supervisor, Geir-Arne Fuglstad, for the invaluable guidance and assistance he has provided throughout the last year. Without his astute and detailed feedback, this report would be far less readable.

Mathias Leander Isaksen,
Trondheim, June 2020.

Abstract

In this thesis, we consider and compare two approaches for non-stationary spatial modeling with Gaussian random fields (GRFs). The first is based on a stochastic partial differential equation (SPDE) with a GRF as its solution. Through discretization, a computationally efficient Gaussian Markov random field (GMRF) approximation is obtained. The non-stationary covariance structure is controlled through the spatially varying coefficients of the SPDE. The R package `R-INLA`, which is an implementation of the Integrated Nested Laplace Approximation (INLA) framework, is used for performing inference and prediction. In the second approach, the GRF is expressed as a convolution between spatially varying kernel functions and Gaussian white noise, so that the non-stationary covariance function is specified indirectly through the kernel functions. For this approach we use the R package `BayesNSGP`, which is dedicated to non-stationary modeling with the kernel-based method.

The SPDE- and kernel-based approaches are presented, and we describe stationary and non-stationary parametrizations. The non-stationarity is modeled through regression on spatial covariates, resulting in an inflexible specification of the covariance structure. The non-stationary SPDE model is implemented in `R-INLA`, along with the parametrizations. The parametrizations used for the kernel-based approach are not available in `BayesNSGP`, and are implemented manually. We compare the two approaches both qualitatively and quantitatively. The inferential properties and predictive power of the models are first investigated through a simulation study, where the observed data is generated from a known process. Afterwards, a case study is carried out with precipitation data from the contiguous United States (CONUS).

The results from the simulation study indicate that there are situations where one approach is more appropriate than the other. In particular, we consider a situation were the approaches lead to qualitatively different covariance structures, and demonstrate that the approach with the correct one performs better. Further, when the observed data is generated from a non-stationary process, the stationary models perform considerably worse than the non-stationary models. In the case study, however, the differences between the stationary and non-stationary models are less dramatic. The SPDE-based models lead to marginally better results, and have considerably faster run-times than the corresponding kernel-based models. While these results indicate that the SPDE-based approach should be preferred, we suggest that more investigation is necessary before any reliable conclusions can be made.

Sammendrag

I denne oppgaven tar vi for oss to tilnærminger til ikke-stasjonær romlig modellering med gaussiske stokastiske felt (GRF-er), og sammenligner disse. Den første tilnærmingen tar i bruk en stokastisk partiell differensialligning (SPDE), som har en GRF som løsning. Ved å diskretisere likningen, får vi en beregningsmessig effektiv Gaussian Markov random field-tilnærming (GMRF). Den ikke-stasjonære kovariansstrukturen kontrolleres gjennom de romlig varierende koeffisientene til SPDE-en. Inferens og prediksjon utføres med R-pakka `R-INLA`, som er en implementasjon av Integrated Nested Laplace Approximation-rammeverket (INLA). I den andre tilnærmingen uttrykkes GRF-en som en konvolusjon mellom romlig varierende kernelfunksjoner og gaussisk hvit støy, som fører til at den ikke-stasjonære kovariansfunksjonen er indirekte spesifisert gjennom kernelfunksjonene. Her bruker vi R-pakka `BayesNSGP`, som er dedikert til ikke-stasjonær modellering med den kernel-baserte tilnærmingen.

De SPDE- og kernel-baserte tilnærmingene presenteres, og vi beskriver stasjonære og ikke-stasjonære parametriseringer. Ikke-stasjonæriteten modelleres gjennom regresjon på romlige kovariater, som fører til en lite fleksibel spesifisering av kovariansstrukturen. Den ikke-stasjonære SPDE-modellen og tilhørende parametriseringer implementeres i `R-INLA`. Parametriseringene som beskrives for den kernel-baserte tilnærmingen er ikke tilgjengelige i `BayesNSGP`, og implementeres manuelt. De to tilnærmingene sammenlignes både kvalitativt og kvantitativt. Modellenes inferensegenskaper og prediktive evner undersøkes først gjennom et simulasjonsstudie, hvor de observerte dataene genereres fra en kjent prosess. Deretter utfører vi et casestudie, hvor vi tar for oss nedbørsdata fra det kontinentale USA (CONUS).

Resultatene fra simulasjonsstudiet indikerer at det finnes situasjoner hvor den ene tilnærmingen er mer egnet enn den andre. Vi tar for oss en situasjon hvor de to tilnærmingene fører til kvalitativt forskjellige kovariansstrukturer, og demonstrerer at tilnærmingen med den riktige strukturen fører til bedre resultater. Videre har de stasjonære modellene betraktelig dårligere resultater enn de ikke-stasjonære, når de observerte dataene genereres fra en ikke-stasjonær prosess. I casestudiet er derimot forskjellene mellom de stasjonære og ikke-stasjonære modellene mindre. De SPDE-baserte modellene fører til marginalt bedre resultater, og har vesentlig kortere kjøretid enn de tilsvarende kernel-baserte modellene. Mens disse resultatene indikerer at den SPDE-baserte tilnærmingen bør foretrekkes, foreslår vi at det må utføres mer utforskning før pålitelige konklusjoner kan trekkes.

Contents

Notation

In Table 1, we list notation and abbreviations used throughout the thesis.

Table 1: Notation and abbreviations used in thesis.

Notation or abbreviation	Meaning
$f(\cdot)$	Function of single variable or random field, depending on context
$K(\cdot,\cdot)$	Function of two variables, also referred to as a kernel
$\boldsymbol{u}$	n-dimensional vector
$\mathbf{A}$	Matrix of dimension $m \times n$
$\mathbf{I}_n$	Identity matrix of dimension $n \times n$
$\mathbf{1}_n$	n-dimensional vector of ones
$\boldsymbol{\Sigma}$	Covariance matrix/kernel matrix
$\mathbf{Q}$	Precision matrix
$\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	n-dimensional Gaussian distribution with expected value $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$p(\cdot)$	A probability density function
$p(\cdot \mid \cdot)$	A conditional probability density function
$\mathbb{1}(\cdot)$	The indicator function, which is equal to 1 when the argument is true and 0 otherwise
GRF	Gaussian random field
GMRF	Gaussian Markov random field
CRPS	Continuous ranked probability scoring
RMSE	Root mean square error
PDF	Probability density function
CDF	Cumulative distribution function

Let $\boldsymbol{u} = (u_1, \dots, u_n) \in \mathbb{R}^n$ be a vector and A a subset of $\{1, \dots, n\}$. Then, the vector $\boldsymbol{u}_A$ is consists of the elements $\{u_i : i \in A\}$ in the same order as the original vector. Similarly, $\boldsymbol{u}_{-A}$ is the vector with the elements given by A removed. We define the shorthand $\boldsymbol{u}_{a:b} = \boldsymbol{u}_{\{a,\dots,b\}}$ for $b \geq a$.

Chapter 1

Introduction

The field of spatial statistics is concerned with the modeling of processes defined over a region, typically in two or three dimensions. Examples of interesting spatial processes are precipitation, temperature, and pressure. Finding accurate models for these processes is crucial for many applications, which include predicting and assessing climate change, and estimating the future output of hydropower plants. It is important to ensure that the model is able to account for the spatial dependencies inherent to the process of interest. This is typically done specifying a model that includes a *Gaussian random field* (GRF) component, where the residual dependency between any two locations is controlled through a covariance function.

The GRFs applied in practice have mostly been limited to stationary covariance functions, which depends only on the relative position of any two locations. As a result, the covariance structure does not vary over the region of interest. This is a strong assumption, which does not necessarily hold when considering real data (Fuglstad et al., 2015b). While there is no such thing as a "true model" when dealing with real data, some models can be considered more correct than others. By letting the covariance structure vary spatially, we can, for example, model processes where the range is spatially varying throughout the region, and obtain correlation structures with varying sizes and shapes. In this way, we obtain a more flexible model that can potentially account for the non-stationary present in many spatial processes. At the same time, non-stationary modeling is not straight-forward. Specifying a non-stationary covariance function is challenging, and the resulting model is more computationally expensive to estimate. With the advent of powerful computers, the latter has become less of an issue, and multiple approaches to non-stationary modeling have been proposed.

In Sampson and Guttorp (1992), a deformation-based method is introduced.

By deforming the region of interest and describing a stationary, isotropic GRF on the deformed region, the GRF becomes non-stationary on the original region. In this way, the method avoids the difficulty of specifying a valid non-stationary covariance function. However, it requires that the process of interest has been observed repeatedly, i.e., that multiple realizations are available. The kernel convolution-based method for Gaussian modeling was first described in Higdon et al. (1999). Essentially, any GRF can be expressed as a convolution between a Gaussian white-noise process and a kernel function. If the shape of the kernel varies over the region, then the resulting GRF becomes non-stationary. This approach is also considered in Paciorek and Schervish (2006), where a convenient formulation of the closed form covariance function is derived. Instead of specifying the covariance function directly, the covariance structure is indirectly determined by the kernel function. Lindgren et al. (2011) consider a stochastic partial differential equation (SPDE) known to have a certain stationary GRF as its solution. By discretizing the SPDE, a Gaussian Markov random field (GMRF) approximation is obtained. This approximation has nice computational properties, and leads to a direct construction of the inverse of the covariance matrix. By letting the coefficients of the SPDE vary spatially, the solution becomes non-stationary. Extensions are described in Fuglstad et al. (2015b,a), where the shape of the covariance structure is allowed to vary according to a vector field. The approach can also be used for modeling processes defined on a sphere, by replacing the Euclidean distance with a metric tensor. This is done in Fuglstad and Castruccio (2020), where SPDEs are used for compressing large climate simulation models.

This thesis builds upon the work done in Isaksen (2019), which focuses solely on non-stationary modeling with the SPDE approach. In this thesis, we focus on both the SPDE- and kernel-based approaches, which lead to a local and global specification of the covariance structure, respectively. The purpose of this thesis is first and foremost to compare the approaches, and to investigate whether there are situations where they lead to significantly different results. In addition, we are interested in comparing the stationary and non-stationary models. This is done both by conducting a simulation study, where both the spatial process and the parameters controlling it are known in advance. A case study is also presented, where precipitation data from the contiguous United States is considered. For both studies, we compare the predictive performance of the models, and the estimated covariance structures. Model inference and prediction is performed in a Bayesian framework. For the SPDE-based models, inference is done with the R package `R-INLA`, which is an implementation of the Integrated Nested Laplace Approximations (INLA) methodology (Rue et al., 2009). The package `BayesNSGP` (Risser and Turek, 2019) is dedicated to non-stationary modeling in the kernel-based approach, and will therefore be used for kernel-based models.

Non-stationary covariance structures can be modeled in a number of ways. Two of the most popular approaches are to specify the functions describing the non-stationarity either through regression on spatial covariates, or by a basis function representation. The latter leads to a very flexible model, and allows for the estimation of general covariance structures. At the same time, such a representation requires many parameters, and leads to models that tend to capture non-existent patterns in the data. The covariate-based representation, however, is far more rigid, and can be described using comparatively few parameters. We have chosen to model the non-stationarity through regression on covariates. Note that the covariate-based parametrizations used in this thesis are not directly available in the aforementioned tools, and had to be implemented manually. In order to do this, we first had to familiarize ourselves with the more technical aspects of both `R-INLA` and `BayesNSGP`.

The thesis is structured as follows: Chapter 2 gives a brief review of the prerequisite material needed for the rest of the thesis. Chapter 3 introduces the SPDE- and kernel-based approaches to non-stationary modeling, and describes the covariate-based parametrizations. In Chapter 4 we define the model used for inference and prediction in the subsequent chapters, followed by a description of the computational tools utilized. Chapter 5 focuses on the simulation study, while Chapter 6 considers the case study where the models are applied to real precipitation data. The thesis concludes with a combined discussion and conclusion in Chapter 7.

Chapter 2

Background

In this chapter we cover the preliminary theory needed later in the thesis. We start by introducing an important class of stochastic processes called Gaussian random fields (GRFs). Next, two approaches for efficient computations with GRFs are described, namely Gaussian Markov random fields (GMRFs) and Vecchia approximations. Finally, we present the scoring rules used for comparing the predictive performance of different models.

2.1 Gaussian random fields (GRFs)

A key characteristic of spatial processes such as surface temperature and atmospheric pressure is that the value of the process tends to be more similar in nearby locations than locations that are far apart. In the modeling of such processes, it is therefore crucial to capture the dependencies between nearby locations. When working in a regression framework, this is often done by specifying a model that contains a random field component. Random fields can be defined in several ways. For our purposes, it is a stochastic process $\{u(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ where the index set $\mathcal{D}$ is a region of Euclidean space, i.e., $\mathcal{D} \subset \mathbb{R}^d$ with $d \geq 1$.

Assume that we are given m observations $y_1, \ldots, y_m$ of a spatial process, observed at corresponding locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m$. A typical model is obtained by assuming that the observed value can be decomposed as $y_i = \eta(\boldsymbol{s}_i) + \varepsilon_i$ for $i = 1, \ldots, m$, where $\varepsilon_1, \ldots, \varepsilon_m \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ are the measurement errors with variance $\sigma_\varepsilon^2 \geq 0$ and $\eta(\boldsymbol{s}_i)$ is the linear predictor

$$\eta(\boldsymbol{s}) = \mu + \boldsymbol{x}(\boldsymbol{s})^\mathsf{T}\boldsymbol{\beta} + u(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D},$$

evaluated in $\boldsymbol{s}_i$. Here $\mu \in \mathbb{R}$ is the intercept, $\boldsymbol{x}(\cdot)$ is a p-dimensional vector-valued function providing covariates, and $\boldsymbol{\beta} \in \mathbb{R}^p$ quantifies the fixed effect of the covariates. The last component, $u(\cdot)$, is a random field intended to capture the residual spatial dependencies not explained by the spatial covariates. Due to their theoretical and computational properties, GRFs are the most common choice for $u(\cdot)$ in such models. Many processes can be argued to be approximately Gaussian through the central limit theorem, making GRFs a convenient choice for modeling. Further, most essential computations involving GRFs, including prediction, reduce to simple linear algebra. A review of GRFs is found in Abrahamsen (1997), which is used as a reference for most of this section. We define GRFs by considering finite-dimensional joint distributions.

Definition 2.1 (Gaussian random field (GRF))**.** Let $\mathcal{D} \in \mathbb{R}^d$ for $d \geq 1$. A random field $u(\cdot) = \{u(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{D}\}$ is said to be a *Gaussian random field* if $(u(\boldsymbol{x}_1), \ldots, u(\boldsymbol{x}_m))$ follows a multivariate Gaussian distribution for any configuration of points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathcal{D}$ for any $m \geq 1$.

We only work with GRFs on regions in Euclidean space, and therefore limit our definition accordingly. However, it is also possible to define GRFs on more general topological spaces, such as manifolds (Adler, 2004).

Let $u(\cdot)$ be a GRF on some region $\mathcal{D} \subset \mathbb{R}^d$ with $d \geq 1$. $u(\cdot)$ is then fully specified by two components. The first is the *mean function* $\mu(\cdot) : \mathcal{D} \to \mathbb{R}$, defined by $\mu(\boldsymbol{x}) = \mathrm{E}[u(\boldsymbol{x})]$ for $\boldsymbol{x} \in \mathcal{D}$. The second is the *covariance function* $C(\cdot,\cdot) : \mathcal{D}^2 \to \mathbb{R}$, which is defined by $C(\boldsymbol{x},\boldsymbol{y}) = \mathrm{Cov}(u(\boldsymbol{x}), u(\boldsymbol{y}))$ for $\boldsymbol{x},\boldsymbol{y} \in \mathcal{D}$. From $C(\cdot,\cdot)$ we can define the *marginal standard deviation function* $\sigma(\cdot) : \mathcal{D} \to \mathbb{R}_\oplus$ given by $\sigma(\boldsymbol{x}) = \sqrt{C(\boldsymbol{x},\boldsymbol{x})}$ for $\boldsymbol{x} \in \mathcal{D}$. Combining these, we obtain the *correlation function* $R(\cdot,\cdot) : \mathcal{D}^2 \to [-1, 1]$:

$$R(\boldsymbol{x},\boldsymbol{y}) = \frac{C(\boldsymbol{x},\boldsymbol{y})}{\sigma(\boldsymbol{x})\sigma(\boldsymbol{y})}, \quad \boldsymbol{x},\boldsymbol{y} \in \mathcal{D}.$$

While both $C(\cdot,\cdot)$ and $R(\cdot,\cdot)$ quantify the amount of dependence between the value of the GRF in any two locations, $R(\cdot,\cdot)$ always takes on a value between -1 and 1, and does not depend on the marginal variance at each location. In general, a correlation close to 1 in absolute value leads to a strong dependency, while correlations close to 0 indicate independence. Note that we can write $C(\boldsymbol{x},\boldsymbol{y}) = \sigma(\boldsymbol{x})\sigma(\boldsymbol{y})R(\boldsymbol{x},\boldsymbol{y})$ for $\boldsymbol{x},\boldsymbol{y} \in \mathcal{D}$, which allows us to describe the GRF through $\mu(\cdot)$, $\sigma(\cdot)$, and $R(\cdot,\cdot)$. In this way, $\mu(\cdot)$ and $\sigma(\cdot)$ determine the distribution of $u(\boldsymbol{s})$ for any $\boldsymbol{s} \in \mathcal{D}$, while $R(\cdot,\cdot)$ alone determines the strength of the dependency between $u(\boldsymbol{x})$ and $u(\boldsymbol{y})$ for $\boldsymbol{x},\boldsymbol{y} \in \mathcal{D}$.

The GRF $u(\cdot)$ is said to be *stationary* if $\mu(\cdot)$ is a constant function ($\mu(\boldsymbol{x}) = \mu_0$ for all $\boldsymbol{x} \in \mathcal{D}$) and the covariance $C(\boldsymbol{x},\boldsymbol{y})$ between any two locations $\boldsymbol{x},\boldsymbol{y} \in \mathcal{D}$ depends only on $\boldsymbol{y} - \boldsymbol{x}$. A GRF with a mean function identically equal to 0 is said to be *centered.* If $u(\cdot)$ has the mean function $\mu(\cdot)$, then the GRF $w(\cdot)$ defined by $w(\boldsymbol{x}) = u(\boldsymbol{x}) - \mu(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathcal{D}$ is centered. Therefore, we can write $u(\boldsymbol{x}) =$

$w(\boldsymbol{x}) + \mu(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{D}$. In other words, any GRF can be written as a sum of a centered GRF and a deterministic function. Since the mean structure can be decoupled in this fashion, our focus is on centered GRFs and their covariance structures.

By sampling $u(\cdot)$ at a finite set of locations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathcal{D}$ with $m \geq 1$, we obtain the vector $\boldsymbol{u} = (u(\boldsymbol{x}_1), \ldots, u(\boldsymbol{x}_m))$. Since $\boldsymbol{u}$ is jointly Gaussian, any linear combination $\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{u}$ with $\boldsymbol{\alpha} \in \mathbb{R}^m$ is also Gaussian. This implies that $\operatorname{Var}(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{u}) = \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\alpha} \geq 0$, where $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{u}$. The covariance function $C(\cdot,\cdot)$ must, therefore, be what is called a positive definite function.

Definition 2.2. Let $\mathcal{D} \subset \mathbb{R}^d$. A kernel $K(\cdot,\cdot) : \mathcal{D}^2 \to \mathbb{R}$ or function $f(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is said to be *positive definite* if

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0 \quad \text{or} \quad \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j f(\boldsymbol{x}_i - \boldsymbol{x}_j) \geq 0,$$

for any configuration of points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathcal{D}$ and weights $(\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^n$ for any $m \geq 1$.

Note that this definition has an inclusive inequality. This differs from the definition of a positive definite matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, where we require that $\boldsymbol{x}^\mathsf{T}\mathbf{A}\boldsymbol{x} > 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$ with $\|\boldsymbol{x}\| \neq 0$.

For what follows, $c(\cdot) : [0, \infty) \to \mathbb{R}$ is a positive definite function. A stationary covariance function $C(\cdot,\cdot)$ is said to be *isotropic* if it only depends on the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$. We can then write $C(\boldsymbol{x}, \boldsymbol{y}) = c(\|\boldsymbol{y} - \boldsymbol{x}\|)$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$. Any stationary $C(\cdot,\cdot)$ that can not be expressed on this form is said to be *anisotropic*. Anisotropy is usually divided into two categories: zonal and geometric. We consider only the latter.

In order to define geometric anisotropy, we introduce a modification of the Euclidean norm. If $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and $\mathbf{S}$ is a $d \times d$ positive definite matrix, then the *Mahalanobis distance* between $\boldsymbol{x}$ and $\boldsymbol{y}$ with respect to $\mathbf{S}$ is

$$h(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^\mathsf{T}\mathbf{S}^{-1}(\boldsymbol{x} - \boldsymbol{y})}.$$

We define $C(\cdot,\cdot)$ to be a *geometrically anisotropic* covariance function if it can be expressed as $C(\boldsymbol{x}, \boldsymbol{y}) = c(h(\boldsymbol{x}, \boldsymbol{y}))$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$, where $h(\cdot,\cdot)$ is a Mahalanobis distance function. For the special case where $\mathbf{S}$ is a multiple of the identity matrix, $h(\cdot,\cdot)$ is proportional to the Euclidean norm and $C(\cdot,\cdot)$ is isotropic. Otherwise, the covariance between $u(\boldsymbol{x})$ and $u(\boldsymbol{y})$ depends not only on the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$, but also on the direction of $\boldsymbol{y} - \boldsymbol{x}$. A covariance function that cannot be expressed as a function of $\boldsymbol{y} - \boldsymbol{x}$ is said to be *non-stationary*.

When $\mathcal{D} \subset \mathbb{R}^2$, the *isocovariance curve* of $C(\cdot,\cdot)$ with respect to some point $\boldsymbol{x} \in \mathcal{D}$ and level α, is the set of points $\boldsymbol{y} \in \mathcal{D}$ with $C(\boldsymbol{x},\boldsymbol{y}) = \alpha$. These isocovariance curves correspond to the level curves of the map $\boldsymbol{y} \mapsto C(\boldsymbol{x},\boldsymbol{y})$ for $\boldsymbol{y} \in \mathcal{D}$. When $C(\cdot,\cdot)$ is geometrically anisotropic, these curves have a known shape.

Theorem 2.1. *Let $\mathcal{D} \subset \mathbb{R}^2$ and $\boldsymbol{x} \in \mathcal{D}$ be fixed. If $C(\cdot,\cdot) : \mathcal{D}^2 \to \mathbb{R}$ is a geometrically anisotropic covariance function, then the level curves of the map $\gamma(\cdot) : \boldsymbol{y} \mapsto C(\boldsymbol{x},\boldsymbol{y})$ for $\boldsymbol{y} \in \mathcal{D}$ are ellipses centered in $\boldsymbol{x}$.*

Proof. Define $C(\boldsymbol{x},\boldsymbol{y}) = c(h(\boldsymbol{x},\boldsymbol{y}))$ as above, and let $\boldsymbol{\tau} = \boldsymbol{y} - \boldsymbol{x}$ for some $\boldsymbol{y} \in \mathcal{D}$. We can then write $h(\boldsymbol{x},\boldsymbol{y})^2 = \boldsymbol{\tau}^\mathsf{T}\mathbf{S}^{-1}\boldsymbol{\tau}$. Let $(\lambda_1, \boldsymbol{v}_1)$ and $(\lambda_2, \boldsymbol{v}_2)$ be the eigenpairs of $\mathbf{S}$, with $\lambda_1 \geq \lambda_2$ and $\|\boldsymbol{v}_1\| = \|\boldsymbol{v}_2\| = 1$. Since $\mathbf{S}$ is positive definite, the eigenvectors form an orthonormal basis for $\mathbb{R}^2$. While $\mathbf{S}$ and $\mathbf{S}^{-1}$ have the same eigenvectors, the eigenvalue of $\mathbf{S}^{-1}$ corresponding to $\boldsymbol{v}_i$ is $1/\lambda_i$. Using this, we can decompose $\boldsymbol{\tau}$ as $\boldsymbol{\tau} = \alpha_1\boldsymbol{v}_1 + \alpha_2\boldsymbol{v}_2$, which leads to

$$\boldsymbol{\tau}^\mathsf{T}\mathbf{S}^{-1}\boldsymbol{\tau} = \boldsymbol{\tau}^\mathsf{T}(\alpha_1\boldsymbol{v}_1/\lambda_1 + \alpha_2\boldsymbol{v}_2/\lambda_2) = \frac{\alpha_1^2}{\lambda_1} + \frac{\alpha_2^2}{\lambda_2}.$$

The equation $h(\boldsymbol{x},\boldsymbol{y}) = c$ for $c > 0$, which corresponds to a level curve of $\gamma(\cdot)$, then defines an ellipse in the coordinate system with $\boldsymbol{x}$ as origin and axes given by the unit vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. In general, $\lambda_1 > \lambda_2$ produces an ellipse with major axis of length $c\sqrt{\lambda_1}$ along $\boldsymbol{v}_1$ and minor axis of length $c\sqrt{\lambda_2}$ along $\boldsymbol{v}_2$, while the case $\lambda_1 = \lambda_2 = \lambda$ reduces to a circle with radius $c\sqrt{\lambda}$. □

An *isocorrelation curve* is defined in an analogous way, using the correlation function instead of the covariance function. Note that for a stationary covariance function $C(\cdot,\cdot)$, we have $C(\boldsymbol{x},\boldsymbol{y}) = \sigma^2 R(\boldsymbol{x},\boldsymbol{y})$ for any $\boldsymbol{x},\boldsymbol{y} \in \mathcal{D}$, where $\sigma^2 > 0$ is the marginal variance and $R(\cdot,\cdot)$ is the correlation function. The isocovariance curve at level α is the same as the isocorrelation curve at level α/σ^2.

In the isotropic and geometrically anisotropic cases, there are many valid choices for the positive definite function $c(\cdot)$. Among the most popular is the *Matérn* covariance function, which is defined as

$$c(h) = \mathcal{M}_\nu(h) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}}\left(\frac{h}{\phi}\right)^\nu K_\nu\left(\frac{h}{\phi}\right), \quad h \geq 0, \tag{2.1}$$

where h is the distance, $\sigma^2 > 0$ is the marginal variance, $\nu > 0$ specifies the smoothness, and $\phi > 0$ controls the range. $K_\nu(\cdot)$ is the modified Bessel function of second kind, order ν. Smoothness in this case refers to the differentiability of realizations in the mean square sense. A smoothness ν results in a GRF with realizations that are $\lceil \nu \rceil - 1$ times differentiable. The range parameter ϕ determines the "range" of the GRF, i.e., the distance at which two locations become practically independent. A useful quantity is the *effective range*, which is given by the empirically obtained relation $\rho = \sqrt{8\nu}\phi$. For $\nu \geq 1/2$, ρ is the distance at which the correlation is approximately 0.14 (Lindgren et al., 2011).

Two often used covariance functions are contained in the Matérn class. When $\nu = 1/2$, the function simplifies to $c(h) = \sigma^2 \exp(-h/\phi)$, which is the *exponential* covariance function. The resulting GRF is continuous but not differentiable, leading to realizations that are non-smooth in nature. As $\nu \to \infty$ we have that $c(h) \to \sigma^2 \exp\left(-h^2/(2\phi^2)\right)$, which is the *Gaussian* covariance function. In contrast to the exponential, the obtained GRF is infinitely differentiable, which results in smooth realizations.

In Figure 2.1, we show three covariance functions, namely the exponential, Gaussian, and Matérn with $\nu = 1$. For each function the marginal variance σ^2 is 1, and the range ϕ is chosen so that the correlation is 0.14 at a distance of 0.5. We define a one-dimensional centered GRF $u(\cdot) = \{u(s) : s \in [0, 1]\}$ based on each covariance function, and generate a single realization. The realization from the Gaussian covariance function is very smooth, while the exponential leads to a realization that is jagged and seemingly non-differentiable. The realization from the Matérn with $\nu = 1$ is somewhere in between, as it is less jagged than the exponential, but not as smooth as the Gaussian.

While most operations involving GRFs reduce to linear algebra and are easy to perform in theory, both the computation time and storage space complexity of these computations quickly become intractable. For example, if $\boldsymbol{u} = (u(\boldsymbol{x}_1), \ldots, u(\boldsymbol{x}_m))$ is the value of some GRF $u(\cdot) = \{u(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{D}\}$ sampled in m locations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathcal{D}$, then the probability density function (PDF) of $\boldsymbol{u}$ is given by

$$p(\boldsymbol{u}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{u} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{u} - \boldsymbol{\mu})\right], \quad \boldsymbol{u} \in \mathbb{R}^n, \tag{2.2}$$

which is the density function of the n-dimensional multivariate Gaussian distribution with expected value $\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{u}]$ and covariance matrix $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{u})$. As $\boldsymbol{\Sigma}$ is an $n \times n$ matrix, the computation time for constructing it and the space needed to store it are both $\mathcal{O}(n^2)$. Computing the inverse $\boldsymbol{\Sigma}^{-1}$ and determinant $|\boldsymbol{\Sigma}|$ is even more expensive, as both operations have a computational complexity of $\mathcal{O}(n^3)$. If $\boldsymbol{\Sigma}^{-1}$ has been computed, then predicting the value of $u(\cdot)$ in k unobserved locations has a computational complexity of $\mathcal{O}(kn^2 + nk^2)$.

As a result, computations with GRFs become prohibitively expensive for large n, leading to the *big n problem*. In the following two sections, we discuss two approximations that reduce both the computational costs and storage needs associated with GRFs.

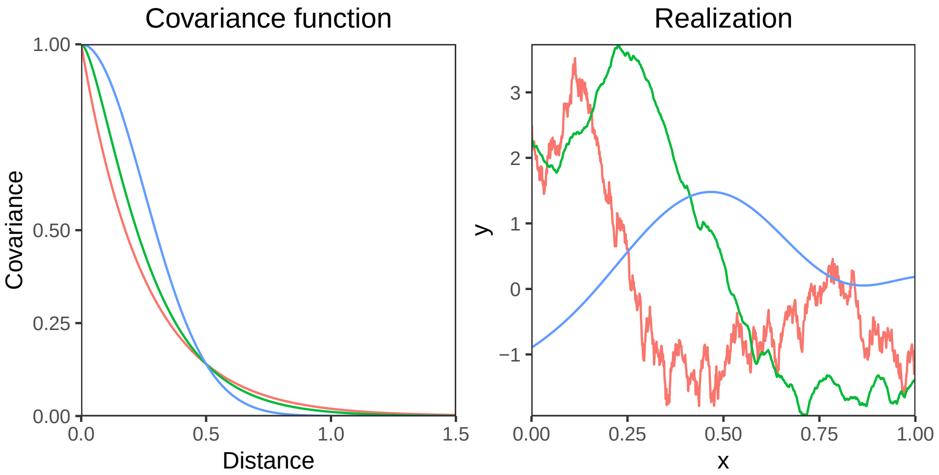


Figure 2.1: The left plot shows the Matérn covariance function for $\nu = 0.5$ (—), $\nu = 1$ (—), and $\nu = \infty$ (—). The marginal variance σ^2 is 1, and the range ϕ is chosen so that the correlation is 0.14 for a distance of 0.5. Based on each of the covariance functions, we define a one-dimensional centered GRF on the interval $[0, 1]$. In the right plot we show a realization from each of the GRFs, using a regular grid of size 1000.

2.2 Gaussian Markov random fields (GMRFs)

The PDF of the multivariate Gaussian distribution in Equation (2.2) depends on the inverse of the covariance matrix, $\mathbf{\Sigma}^{-1}$. This matrix is referred to as the *precision matrix*, and is usually denoted by $\mathbf{Q}$. The precision matrix and its properties motivate the definition of *Gaussian Markov random fields* (GMRFs), which are described in this section. A comprehensive description of GMRFs and their applications is found in Rue and Held (2005). Before defining GMRFs, we introduce two necessary concepts.

Let X, Y and Z be random variables. X and Y are said to be *conditionally independent* given Z if $p(x, y|z) = p(x|z)p(y|z)$, where $p(\cdot|\cdot)$ is the conditional probability density function of its arguments. This is denoted by $X \perp Y \mid Z$. The following theorem connects conditional independence to the precision matrix.

Theorem 2.2. *Let $\boldsymbol{v} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{Q}^{-1})$ and $\boldsymbol{v}_{-ij}$ be $\boldsymbol{v}$ with the elements at indices i and j removed. Then, for $i \neq j$, v_i is conditionally independent of v_j given $\boldsymbol{v}_{-ij}$ if and only if* $\mathrm{Q}_{ij} = 0$,

$$\mathrm{Q}_{ij} = 0 \Longleftrightarrow v_i \perp v_j \mid \boldsymbol{v}_{-ij}.$$

Proof. See Section 2.2 in Rue and Held (2005). □

A *labeled graph* $\mathcal{G}$ consists of the pair $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, \ldots, n\}$ is the set of vertices and $\mathcal{E}$ is the set of edges. In an undirected graph, elements $i, j \in \mathcal{V}$ are connected to each other if $\{i, j\} \in \mathcal{E}$. Since we are only going to use undirected graphs, the term "graph" means "undirected graph" for the rest of the thesis. The defintion of GMRFs links the conditional independence structure of $\mathbf{Q}$ to a labeled graph $\mathcal{G}$.

Definition 2.3 (Gaussian Markov random field (GMRF))**.** Let $\boldsymbol{v}$ be an n-dimensional Gaussian vector with mean vector $\boldsymbol{\mu}$ and precision matrix $\mathbf{Q}$, and let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a labeled graph with vertices $\mathcal{V} = \{1, \ldots, n\}$ and edges $\mathcal{E}$. Then $\boldsymbol{v}$ is said to be a *Gaussian Markov random field* with respect to $\mathcal{G}$ if

$$\mathrm{Q}_{ij} \neq 0 \Longleftrightarrow \{i, j\} \in \mathcal{E}$$

for all $i \neq j$. The PDF of $\boldsymbol{v}$ is

$$p(\boldsymbol{v}) = \frac{|\mathbf{Q}|^{1/2}}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}(\boldsymbol{v} - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{Q}(\boldsymbol{v} - \boldsymbol{\mu})\right], \quad \boldsymbol{v} \in \mathbb{R}^n.$$

The graph $\mathcal{G}$ reflects the conditional independence structure of the GMRF. The condition

$$v_i \perp v_j \mid \boldsymbol{v}_{-ij} \Longleftrightarrow \{i, j\} \in \mathcal{V}$$

is called the *pairwise Markov property.* This property is equivalent to two other properties, which are listed in Theorem 2.3. Before stating the theorem, we need to introduce the concept of separating sets. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph, and let $A, B, C \subset \mathcal{V}$ be disjoint sets of vertices. C is said to *separate* A and B if any path from an element in A to an element in B has to visit an element of C.

Theorem 2.3. *An n-dimensional GMRF $\boldsymbol{v}$ with graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ will, in addition to the pairwise Markov property, always satisfy the following properties, which are all equivalent to each other:*

- *The local Markov property: For any $i \in \mathcal{V}$,*

  $$v_i \perp \boldsymbol{v}_{-\{i,ne(i)\}} \mid \boldsymbol{v}_{ne(i)}$$

  where $ne(i)$ are the neighbors of vertex i, i.e., $ne(i) = \{j : \{i, j\} \in \mathcal{E}\}$.

- *The global Markov property: If A, B and C are disjoint subsets of $\mathcal{V}$ such that C separates A and B, then*

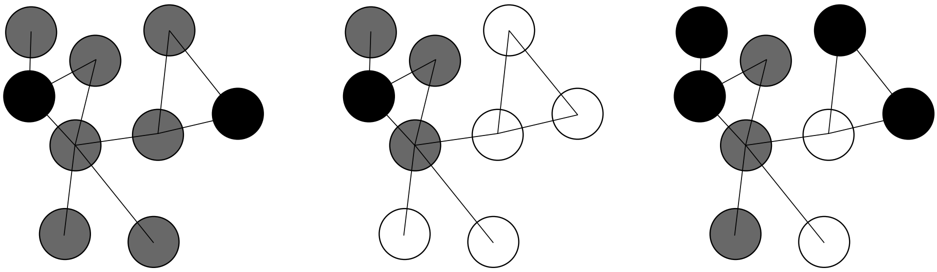
  $$\boldsymbol{v}_A \perp \boldsymbol{v}_B \mid \boldsymbol{v}_C,$$

  as long as both A and B are non-empty.

Proof. See Section 2.2 in Rue and Held (2005). □

The three properties are illustrated in Figure 2.2, where a GMRF $\boldsymbol{v} = (v_1, \ldots, v_9)$ is represented by its graph structure.

Since no limitations are put on the graph $\mathcal{G}$, any Gaussian vector is a GMRF with respect to the graph implied by its precision matrix. However, the benefits of the GMRF formulation are attained when the precision matrix is *sparse.* A matrix is said to be sparse when the number of non-zero elements is small in comparison to the total number of elements. Sparse matrices can be stored by specifying the positions and values of only the non-zero elements, which is considerably cheaper than storing the entire matrix. They also allow for significantly faster computation of many important numerical linear algebra operations, such as solving linear systems and computing the Cholesky decomposition. See Isaksen (2019) for a discussion of these. In general, computing the Cholesky decomposition of an $n \times n$ precision matrix $\mathbf{Q}$ has a time complexity of $\mathcal{O}(n^3)$. For temporal, spatial, and spatio-temporal GMRFs, this is reduced to $\mathcal{O}(n)$, $\mathcal{O}(n^{1.5})$, and $\mathcal{O}(n^2)$, respectively (Rue and Held, 2005, Section 2.3). Note that the connection between $\mathbf{Q}$ and $\mathcal{G}$ allows us to take advantage of theory and algorithms regarding graphs, for computations involving $\mathbf{Q}$.



(a) Pairwise (b) Local (c) Global

Figure 2.2: Illustrations of Markov properties described in Theorem 2.3. Gray nodes represent the elements that are being conditioned upon, while the elements under consideration are black. The remaining elements are shown as white nodes. (a) The pairwise Markov property guarantees that the black nodes are conditionally independent given the gray nodes. (b) By the local Markov property, the black node is conditionally independent of the white nodes given its gray neighbor nodes. (c) Due to the global Markov property, the two groups of black nodes are conditionally independent given the gray nodes, since the gray nodes separate the groups.

From a more practical aspect, the usefulness of GMRFs rests on the precision matrix $\mathbf{Q}$ being sparse and possible to compute in reasonable time. Say, for example, that $\mathbf{Q}$ can only be obtained by inverting the dense covariance matrix $\mathbf{\Sigma}$. Then the computational cost of the inversion becomes a bottleneck. Gaussian models that lead to a closed form specification of a sparse precision matrix are therefore particularly attractive. Examples are auto-regressive (AR) processes, the BYM model (Besag et al., 1991), and the SPDE-based GRF approximation outlined in Lindgren et al. (2011). The latter is described in Chapter 3.

2.3 Vecchia approximations

While GMRFs allow for efficient computation, there are many situations where the precision matrix is not easily obtainable, and we must construct the covariance matrix directly. One way to reduce the computational cost is through something called a *Vecchia approximation*, first described in Vecchia (1988). Before introducing the Vecchia approximation, we give some motivation.

Let $\boldsymbol{y} = (y_1, \ldots, y_m)$ be an n-dimensional Gaussian vector with PDF $p(\cdot)$.

Then, $p(\cdot)$ can always be factored as

$$p(\boldsymbol{y}) = p(y_1) \prod_{i=2}^{m} p(y_i \mid \boldsymbol{y}_{1:(i-1)}).$$

The vector subscript notation is described in Notation, and is used throughout this section. This form does not offer any computational advantages. For each $i = 2, \ldots, m$, obtaining the factor $p(y_i \mid \boldsymbol{y}_{1:(i-1)})$ involves computing the inverse of an $(i-1) \times (i-1)$ matrix, which becomes a bottleneck for large i. We can make this more efficient by, instead of conditioning on all preceding variables, conditioning only on a subset. By replacing the conditioning vector $\boldsymbol{y}_{1:(i-1)}$ with a subvector $\boldsymbol{y}_{q(i)}$ such that $q(i) \subset \{1, \ldots, i-1\}$, the resulting approximation is

$$p(\boldsymbol{y}) \approx \hat{p}(\boldsymbol{y}) = p(y_1) \prod_{i=2}^{m} p(y_i \mid \boldsymbol{y}_{q(i)}). \tag{2.3}$$

A Markov assumption is made, as we assume that y_i is conditionally independent of the preceding elements not in $q(i)$, given those in $q(i)$. The approximation depends both on the ordering of $\boldsymbol{y}$ and how the conditioning sets $q(i)$ are chosen. If we ensure that $|q(i)| \leq k$ for each i, then the computation of this approximation involves inverting matrices of size $k \times k$ or smaller.

A popular way to do this is the AR(k) model. For each i, $q(i)$ is chosen to be the, at most, k indices directly preceding i,

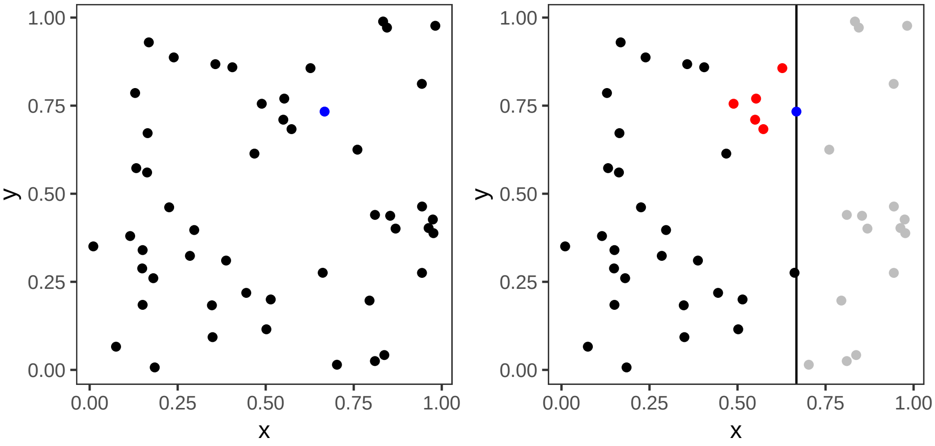
$$q(i) = \begin{cases} \{1, \ldots, i-1\}, & \text{if } i < k, \\ \{i-k, \ldots, i-1\}, & \text{if } i \geq k. \end{cases}$$

Among the previous values, $\boldsymbol{y}_{q(i)}$ consists of the $\min\{i-1, k\}$ that are the closest to y_i in index. This is a sensible approach when the elements of $\boldsymbol{y}$ are, for example, indexed by time and ordered accordingly. However, when dealing with spatially observed data, we can choose the conditioning vector in a more appropriate way.

Vecchia (1988) considers the case where the vector $\boldsymbol{y} = (y_1, \ldots, y_m)$ contains observations of the spatial process $y(\cdot)$ over a region $\mathcal{D} \subset \mathbb{R}^2$, so that $y_i = y(\boldsymbol{s}_i)$ with $\boldsymbol{s}_i \in \mathcal{D}$. This is then modeled as

$$y_i = \mu + \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta} + u_i + \varepsilon_i, \quad i = 1, \ldots, m,$$

where $\mu \in \mathbb{R}$ is the intercept, $\boldsymbol{x}_i \in \mathbb{R}^p$ contains spatial covariates at $\boldsymbol{s}_i$, $\boldsymbol{\beta} \in \mathbb{R}^p$ quantifies the linear effect of the covariates, u_i is the value of a GRF $u(\cdot)$ evaluated at $\boldsymbol{s}_i$, and $\varepsilon_1, \ldots, \varepsilon_m \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ are the measurement errors. In this context, conditioning on all preceding data seems excessive, as the observations made close to y_i are usually much more important for determining its value than those made



(a) (b)

Figure 2.3: (a) 50 locations in $[0,1]^2$. (b) Demonstration of conditioning locations with the Vecchia approximation, using $k = 5$ and ordering by increasing x-coordinate.

far away. Therefore, the conditioning set $q(i)$ is chosen by using the $\min\{i-1, k\}$ earlier indexed observations closest to y_i in location. In other words, the locations $\boldsymbol{s}_j$ for $j \in q(i)$ are those among $\{\boldsymbol{s}_j : j \in \{1, \ldots, i-1\}\}$ that minimize the ordinary Euclidean distance $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$. The approximation in Equation (2.3) also depends on the ordering of $\boldsymbol{y}$. Vecchia (1988) suggests ordering by increasing x- or y-coordinate. This is discussed in Guinness (2018), where more technical ordering schemes are shown to lead to better results.

Figure 2.3a shows an example of 50 observations locations with $\mathcal{D} = [0,1]^2$. We order the locations by increasing x-coordinate, let $k = 5$, and consider the blue location $\boldsymbol{s}$. In Figure 2.3b, the locations preceding $\boldsymbol{s}$ in index are those to the left of the black line. Among these, the 5 closest to $\boldsymbol{s}$ are colored red. The locations to the right of the line cannot be conditioned upon, and are colored grey.

2.4 Scoring rules for predictions

In order to evaluate and compare the predictive performance of different models, we need some way to quantify how good point predictions and predictive distributions are compared to observed values. We are interested in predicting the value of $X \mid$ data, i.e., a random variable X conditioned on observed data. While a point prediction $\hat{x}$ of $X \mid$ data consists of a single value, a predictive distribution density function $\hat{f}(\cdot)$ specifies a predictive probability distribution for $X \mid$ data. This allows for the uncertainty of the prediction to be accounted for.

Point predictions are evaluated using the *root mean square error* (RMSE). Given a vector of predictions $\hat{\boldsymbol{x}} = (\hat{x}_1, \dots, \hat{x}_n) \in \mathbb{R}^n$ and observed values $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$, the RMSE of the predicted values is defined as

$$\text{RMSE}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2}.$$

When comparing two models, the model with the lowest RMSE is preferred. Since the RMSE depends on the square of the deviations, it is sensitive to outliers.

Predictive distributions are evaluated using the *continuous ranked probability score* (CRPS), which is described in Gneiting and Raftery (2007). For a predictive distribution $\hat{f}(\cdot)$ with corresponding CDF $\hat{F}(\cdot)$ and observed value x, the CRPS of the predictive distribution is defined to be

$$\text{CRPS}(\hat{F}(\cdot), x) = \int_{-\infty}^{\infty} (\hat{F}(z) - \mathbb{1}(x \leq z))^2 \mathrm{d}z,$$

where $\mathbb{1}(\cdot)$ is the indicator function: $\mathbb{1}(x \leq z) = 1$ if $x \leq z$ and 0 otherwise. Like the RMSE, a lower value of the CRPS is preferred. For prediction of multiple random variables $\boldsymbol{X} = (X_1 \dots, X_n)$, let $\hat{F}_i(\cdot)$ be the predictive CDF of $X_i \mid$ data and x_i be the observed value. The *mean CRPS* of the predictive distributions is then

$$\overline{\text{CRPS}} = \frac{1}{n}\sum_{i=1}^{n}\text{CRPS}(\hat{F}_i(\cdot), x_i).$$

In the context where a predictive distribution $\hat{f}(\cdot)$ is available and a point prediction is needed, the predictive mean $\hat{x} = \int x\hat{f}(x)\,\mathrm{d}x$ is a common choice.

The CRPS is a proper scoring rule, while the RMSE is not. In this context, this means that the CRPS will, on average, prefer the true model that the data was generated from. The RMSE, on the other hand, will prefer the model that gives predictive means closest to the true observed values. This is illustrated in Figure 2.4. In each plot, a possible predictive distribution density function of $X \mid$ data is plotted in blue. The left density $\hat{f}_1(\cdot)$ is Gaussian with $\mu = 0.9$ and $\sigma = 0.1$, while the right density $\hat{f}_2(\cdot)$ is Gaussian with $\mu = 1$ and $\sigma = 1.2$. The predicted means $\hat{x}_1 = 0.9$ and $\hat{x}_2 = 1$ are shown as dashed red lines. Finally, the observed value $x = 0$ is shown as a solid black line. The RMSEs are $\text{RMSE}_1 = \sqrt{(0.9 - 0)^2} = 0.9$ and $\text{RMSE}_2 = \sqrt{(1 - 0)^2} = 1$, while the CRPS values are $\text{CRPS}_1 = 0.844$ and $\text{CRPS}_2 = 0.595$.

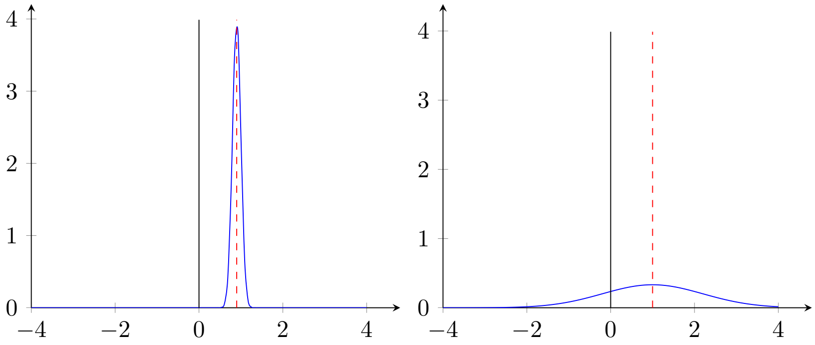


Figure 2.4: Prediction distributions of a random variable $X \mid \text{data}$. Both plots show Gaussian densities (—), with $\mu = 0.9$ and $\sigma = 0.1$ in the left, and $\mu = 1$ and $\sigma = 1.2$ in the right. The predictive means (- - -) and the observed value $x = 0$ (—) are also shown.

Based on the RMSE, the first prediction is better. However, the predictive distribution $\hat{f}_1(\cdot)$ is very sharp and fails to explain the observed value. This is accounted for with the CRPS. While the predictive mean $\hat{x}_2$ misses by more than $\hat{x}_1$, the higher uncertainty of $\hat{f}_2(\cdot)$ captures the observed value, which leads to a lower CRPS.

Chapter 3

Beyond stationarity

When modeling real-life processes, the assumption of stationarity in the covariance function is unrealistic, as it requires that the dependence structure is the same throughout the region of interest. With a non-stationary covariance structure, we can let properties such as range and marginal variance be spatially varying, leading to a model that is more flexible. However, given only a single realization of the observed data, it is, in general, not possible to identify and separate the covariance structure from the mean structure (Gelfand et al., 2010, page 30). As a result, patterns in the data that are best explained by the mean structure, might instead be captured by the covariance structure. It is, nevertheless, possible that models with well-specified non-stationary covariance structures lead to better predictions.

In this chapter, we introduce two approaches for specifying GRFs with non-stationary covariance functions. First, we describe a method based on a *stochastic partial differential equation* (SPDE), where the covariance structure is determined indirectly from spatially varying coefficients. Second, we present a kernel convolution-based method, where the covariance structure is indirectly specified on a closed form by spatially varying kernel functions.

3.1 Stochastic partial differential equations

3.1.1 Specifying covariance structure

SPDEs are partial differential equations characterized by the introduction of stochastic terms and coefficients. Our focus is on the SPDE

$$\left(\kappa^2 - \Delta\right)(\tau u(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathbb{R}^2, \tag{3.1}$$

which was first considered in Whittle (1954), and later connected to GMRFs in Lindgren et al. (2011). Here $\Delta = \nabla \cdot \nabla$ is the Laplace operator and $\kappa, \tau > 0$ are constants. On the right-hand side we have a spatial standard Gaussian white noise process $\mathcal{W}(\cdot)$. It is characterized by $\int_A \mathcal{W}(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s} \sim \mathcal{N}(0, |A|)$ for measurable a $A \subset \mathbb{R}^2$, and the fact that

$$\left(\int_{A_1} \mathcal{W}(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s}, \ldots, \int_{A_n} \mathcal{W}(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s}\right)$$

is multivariate Gaussian for measurable $A_1, \ldots, A_n \subset \mathbb{R}^2$ with $n \geq 1$. The stationary solution of SPDE (3.1) is a GRF with a Matérn covariance function,

$$C(\boldsymbol{s}_1, \boldsymbol{s}_2) = \mathrm{Cov}(u(\boldsymbol{s}_1), u(\boldsymbol{s}_2)) = \frac{1}{4\pi\kappa^2\tau^2}(\kappa\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|)K_1(\kappa\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|), \quad \boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathbb{R}^2.$$

By comparison with Equation (2.1) we see that the smoothness is 1 and the range is $1/\kappa$, leading to an effective range of $\sqrt{8}/\kappa$. The marginal variance is equal to $1/(4\pi\kappa^2\tau^2)$, and decreases with both κ and τ. Further, the covariance function is isotropic, as it only depends on the distance between any two locations.

In Fuglstad et al. (2015a,b), a positive definite 2×2 matrix $\mathbf{H}$ is used to modify the Laplacian, leading to the SPDE

$$\left(\kappa^2 - \nabla \cdot \mathbf{H}\nabla\right)(\tau u(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathbb{R}^2. \tag{3.2}$$

The stationary solution to this SPDE is a GRF where the covariance between the locations $\boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathbb{R}^2$ is given by (Fuglstad et al., 2015b)

$$C(\boldsymbol{s}_1, \boldsymbol{s}_2) = \frac{1}{4\pi\kappa^2\tau^2\,|\mathbf{H}|^{1/2}}(\kappa\|\mathbf{H}^{-1/2}(\boldsymbol{s}_1 - \boldsymbol{s}_2)\|)K_1(\kappa\|\mathbf{H}^{-1/2}(\boldsymbol{s}_1 - \boldsymbol{s}_2)\|). \tag{3.3}$$

Since $\|\mathbf{H}^{-1/2}(\boldsymbol{s}_1 - \boldsymbol{s}_2)\|$ is the Mahalanobis distance between $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ with respect to $\mathbf{H}$, the solution exhibits geometric anisotropy.

Let $(\lambda_1, \boldsymbol{q}_1)$ and $(\lambda_2, \boldsymbol{q}_2)$ be eigenpairs of $\mathbf{H}$ satisfying $\lambda_1 \geq \lambda_2$ and $\|\boldsymbol{q}_1\| = \|\boldsymbol{q}_2\| = 1$. Note that $\mathbf{H}$ being positive definite implies that $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ are orthogonal. We let $\boldsymbol{q}_1$ be in the upper half-plane and $\boldsymbol{q}_2$ be $\boldsymbol{q}_1$ rotated 90 degrees counter-clockwise. The special case $\mathbf{H} = \mathbf{I}_2$ corresponds to SPDE (3.1), where the effective range is $\sqrt{8}/\kappa$ in every direction. By combining this with Theorem 2.1, which tells us that the range along the direction of $\boldsymbol{q}_i$ scales with $\sqrt{\lambda_i}$, we get the longest and shortest effective ranges

$$\rho_1 = \frac{\sqrt{8}}{\kappa}\sqrt{\lambda_1} \text{ and } \rho_2 = \frac{\sqrt{8}}{\kappa}\sqrt{\lambda_2} \tag{3.4}$$

in the direction of $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$, respectively. The *strength of the anisotropy*, which is the ratio between the longest and shortest range, is equal to the ratio $\sqrt{\lambda_1/\lambda_2}$.

From Equation (3.3), we see that the marginal variance of the solution is

$$\sigma^2 = \frac{1}{4\pi\kappa^2\tau^2 \left|\mathbf{H}\right|^{1/2}} = \frac{1}{4\pi\kappa^2\tau^2\sqrt{\lambda_1\lambda_2}}, \tag{3.5}$$

which depends on $\mathbf{H}$ in addition to κ and τ.

If we define $\hat{\tau} = \tau/\kappa^2$ and $\hat{\mathbf{H}} = \mathbf{H}/\kappa^2$, then SPDE (3.2) can be expressed as

$$\left(1 - \nabla\cdot\hat{\mathbf{H}}\nabla\right)(\hat{\tau}u(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s}\in\mathbb{R}^2.$$

Using all three parameters therefore leads to overparametrization, which is avoided by fixing $\kappa = 1$ or $\tau = 1$. In this thesis, we fix the value of κ, so that $\mathbf{H}$ alone determines the correlation structure.

Both of the SPDEs described so far lead to stationary GRFs. In Lindgren et al. (2011) and Fuglstad et al. (2015a) non-stationarity is introduced by letting the coefficients vary spatially. This leads to the SPDE

$$\left(\kappa^2(\boldsymbol{s}) - \nabla\cdot\mathbf{H}(\boldsymbol{s})\nabla\right)(\tau(\boldsymbol{s})u(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s}\in\mathbb{R}^2, \tag{3.6}$$

where, for all $\boldsymbol{s}\in\mathbb{R}^2$, $\mathbf{H}(\boldsymbol{s})$ is positive definite 2×2 matrix, $\kappa(\boldsymbol{s}) > 0$, and $\tau(\boldsymbol{s}) > 0$. For $\boldsymbol{s}\in\mathbb{R}^2$ we define $(\lambda_1(\boldsymbol{s}), \boldsymbol{q}_1(\boldsymbol{s}))$ and $(\lambda_2(\boldsymbol{s}), \boldsymbol{q}_2(\boldsymbol{s}))$ to be the eigenpairs of $\mathbf{H}(\boldsymbol{s})$, with $\lambda_1(\boldsymbol{s}) \geq \lambda_2(\boldsymbol{s})$, $\|\boldsymbol{q}_1(\boldsymbol{s})\| = \|\boldsymbol{q}_2(\boldsymbol{s})\| = 1$, $\boldsymbol{q}_1(\boldsymbol{s})$ in the upper half-plane, and $\boldsymbol{q}_2(\boldsymbol{s})$ obtained by rotating $\boldsymbol{q}_1(\boldsymbol{s})$ 90 degrees counter-clockwise.

The connection between the spatially varying coefficients and the resulting GRF are investigated in Fuglstad et al. (2015a) and Isaksen (2019). Even though Equations (3.4) and (3.5) only hold for constant coefficients, they comply well with the qualitative results: the correlation structure tends to have longer ranges in the direction of $\boldsymbol{q}_1(\cdot)$, and shorter in the direction of $\boldsymbol{q}_2(\cdot)$. In regions where $\kappa(\cdot)$ is large, the range is short, while a smaller $\kappa(\cdot)$ leads to longer ranges. Further, the marginal variance decreases for increasing values of $\kappa(\cdot)$ and $|\mathbf{H}(\cdot)|$.

Based on these observations, for each $\boldsymbol{s}\in\mathbb{R}^2$ we suggest the approximations

$$\rho_1(\boldsymbol{s}) = \frac{\sqrt{8}}{\kappa(\boldsymbol{s})}\sqrt{\lambda_1(\boldsymbol{s})} \text{ and } \rho_2(\boldsymbol{s}) = \frac{\sqrt{8}}{\kappa(\boldsymbol{s})}\sqrt{\lambda_2(\boldsymbol{s})} \tag{3.7}$$

for the effective ranges in direction $\boldsymbol{q}_1(\boldsymbol{s})$ and $\boldsymbol{q}_2(\boldsymbol{s})$, and

$$\tilde{\sigma}^2(\boldsymbol{s}) = \frac{1}{4\pi\kappa^2(\boldsymbol{s})\tau^2(\boldsymbol{s})\left|\mathbf{H}(\boldsymbol{s})\right|^{1/2}} = \frac{1}{4\pi\kappa^2(\boldsymbol{s})\tau^2(\boldsymbol{s})\sqrt{\lambda_1(\boldsymbol{s})\lambda_2(\boldsymbol{s})}} \tag{3.8}$$

for the marginal variance. The function $\rho_2(\cdot)$ is referred to as the baseline effective range function. These approximations are only appropriate when $\kappa(\cdot)$, $\tau(\cdot)$ and the elements of $\mathbf{H}(\cdot)$ are slowly varying functions of $\boldsymbol{s}$. Similar approximations are suggested in Section 6.5 of Blangiardo and Cameletti (2015) and in Marques et al. (2019), where $\kappa(\cdot)$ and $\tau(\cdot)$ are allowed to vary spatially. While we still use the term "range" for non-stationary GRFs, it does not translate directly from the stationary case. For stationary GRFs, the range is a global property that holds over the entire region of interest. In a non-stationary GRF, the dependence structures can vary throughout the region, leading to different ranges depending on both location and direction (Fuglstad et al., 2015b). The functions $\rho_1(\cdot)$ and $\rho_2(\cdot)$ should therefore not be interpreted as direct approximations of the longest and shortest effective ranges in each location, which ρ_1 and ρ_2 represent in the stationary case. However, they give a qualitative idea of how the shape of the dependence structure varies throughout the region.

Analogous to the stationary case, describing SPDE (3.6) using all three of the functions $\mathbf{H}(\cdot)$, $\kappa(\cdot)$, and $\tau(\cdot)$ leads to overparametrization. The overparametrization can be avoided by fixing $\kappa(\boldsymbol{s}) \equiv 1$ or $\tau(\boldsymbol{s}) \equiv 1$ for all $\boldsymbol{s} \in \mathbb{R}^2$. Later, in Section 3.1.5, we describe parametrizations for both alternatives.

3.1.2 Parametrization of $\mathbf{H}(\cdot)$

When specifying $\mathbf{H}(\cdot)$, we need to ensure that both eigenvalues are positive for all $\boldsymbol{s} \in \mathbb{R}^2$. This can be achieved by decomposing $\mathbf{H}(\cdot)$ into two components:

$$\mathbf{H}(\boldsymbol{s}) = \gamma(\boldsymbol{s}) \left(\mathbf{I}_2 + \boldsymbol{w}(\boldsymbol{s})\boldsymbol{w}(\boldsymbol{s})^{\mathsf{T}}\right), \quad \boldsymbol{s} \in \mathbb{R}^2, \tag{3.9}$$

where $\gamma(\boldsymbol{s}) > 0$ for all $\boldsymbol{s} \in \mathbb{R}^2$ and $\boldsymbol{w}(\cdot) = (w_x(\cdot), w_y(\cdot))$ is a vector field. This is similar to the decomposition used in Fuglstad and Castruccio (2020). The eigenvalues of $\mathbf{H}(\boldsymbol{s})$ are $\lambda_1(\boldsymbol{s}) = \gamma(\boldsymbol{s})(1 + \|\boldsymbol{w}(\boldsymbol{s})\|^2)$ and $\lambda_2(\boldsymbol{s}) = \gamma(\boldsymbol{s})$, with corresponding eigenvectors chosen to be $\boldsymbol{q}_1(\boldsymbol{s}) = (w_x(\boldsymbol{s}), w_y(\boldsymbol{s}))$ and $\boldsymbol{q}_2(\boldsymbol{s}) = (-w_y(\boldsymbol{s}), w_x(\boldsymbol{s}))$. The interpretation of each component is best understood by first considering a special case.

When each coefficient is constant, i.e., $\tau(\boldsymbol{s}) \equiv \tau_0$, $\kappa(\boldsymbol{s}) \equiv \kappa_0$, $\gamma(\boldsymbol{s}) \equiv \gamma_0$ and $\boldsymbol{w}(\boldsymbol{s}) \equiv \boldsymbol{w}_0$ for all $\boldsymbol{s} \in \mathbb{R}^2$, SPDE (3.2) with a stationary solution is obtained. The effective ranges and marginal variance are then

$$\rho_1 = \frac{\sqrt{8}}{\kappa_0}\sqrt{\gamma_0(1 + \|\boldsymbol{w}_0\|^2)},\ \rho_2 = \frac{\sqrt{8}}{\kappa_0}\sqrt{\gamma_0},\ \text{and } \sigma^2 = \frac{1}{4\pi\kappa_0^2\tau_0^2\gamma\sqrt{1 + \|\boldsymbol{w}_0\|^2}},$$

and the direction of maximum range is given by $\boldsymbol{w}_0$. We see that γ_0 controls the baseline effective range, i.e., the effective range without any anisotropy present. The strength of the additional anisotropy is specified by $\|\boldsymbol{w}_0\|$.

Based on this, we get a qualitative idea of how the spatially varying coefficients affect the solution. In each location, the scalar function $\gamma(\cdot)$ quantifies the baseline isotropic effect, while $\boldsymbol{w}(\cdot) = (w_x(\cdot), w_y(\cdot))$ specifies the direction and magnitude of the local additional anisotropy. The resulting GRF can be thought of as "different Matérn like fields locally each with its own anisotropy that are combined into a full process" (Fuglstad et al., 2015a). Using this decomposition, the approximate effective range functions are

$$\rho_1(\boldsymbol{s}) = \frac{\sqrt{8}}{\kappa(\boldsymbol{s})}\sqrt{\gamma(\boldsymbol{s})(1+\|\boldsymbol{w}(\boldsymbol{s})\|^2)} \text{ and } \rho_2(\boldsymbol{s}) = \frac{\sqrt{8}}{\kappa(\boldsymbol{s})}\sqrt{\gamma(\boldsymbol{s})},$$

and the approximate marginal variance function is

$$\tilde{\sigma}^2(\boldsymbol{s}) = \frac{1}{4\pi\kappa^2(\boldsymbol{s})\tau^2(\boldsymbol{s})\gamma(\boldsymbol{s})\sqrt{1+\|\boldsymbol{w}(\boldsymbol{s})\|^2}}.$$

In Fuglstad et al. (2015a,b), $\mathbf{H}(\cdot)$ is decomposed as

$$\mathbf{H}(\boldsymbol{s}) = \gamma(\boldsymbol{s})\mathbf{I}_2 + \boldsymbol{v}(\boldsymbol{s})\boldsymbol{v}(\boldsymbol{s})^{\mathsf{T}}, \quad \boldsymbol{s} \in \mathbb{R}^2,$$

where $\boldsymbol{v}(\cdot) = (v_x(\cdot), v_y(\cdot))$. This is equivalent to Equation (3.9) with $\boldsymbol{w}(\cdot) = \boldsymbol{v}(\cdot)/\sqrt{\gamma(\cdot)}$. While $\gamma(\cdot)$ controls the baseline isotropic effect as before, $\boldsymbol{v}(\cdot)$ quantifies the absolute size of the local additional anisotropy. This is opposed to $\boldsymbol{w}(\cdot)$, which specifies the relative size of the additional anisotropy. The eigenvalues of $\mathbf{H}(\boldsymbol{s})$ under this parametrization are $\lambda_1(\boldsymbol{s}) = \gamma(\boldsymbol{s}) + \|\boldsymbol{v}(\boldsymbol{s})\|^2$ and $\lambda_2(\boldsymbol{s}) = \gamma(\boldsymbol{s})$, with accompanying eigenvectors $\boldsymbol{q}_1(\boldsymbol{s}) = (v_x(\boldsymbol{s}), v_y(\boldsymbol{s}))$ and $\boldsymbol{q}_2(\boldsymbol{s}) = (-v_y(\boldsymbol{s}), v_x(\boldsymbol{s}))$.

In the absolute parametrization, the strength of the anisostropy, i.e., the ratio between the longest and shortest range, depends on both $\gamma(\cdot)$ and $\boldsymbol{v}(\cdot)$. With the relative parametrization, it depends only on $\boldsymbol{w}(\cdot)$. Due to this separation, the function $\boldsymbol{w}(\cdot)$, and the effect it has on the SPDE, is easier to interpret than $\boldsymbol{v}(\cdot)$. Therefore, for the rest of this thesis, we use only the relative parametrization.

3.1.3 Role of $\tau(\cdot)$

We illustrate how the function $\tau(\cdot)$ affects the solution. Consider the SPDE

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

which is investigated in Isaksen (2019). It is a special case of SPDE (3.6) with $\tau(\boldsymbol{s}) \equiv 1$ for $\boldsymbol{s} \in \mathbb{R}^2$. However, the solution $u(\cdot)$ to SPDE (3.6) can also be obtained from $\hat{u}(\cdot)$, by defining $u(\cdot) = \hat{u}(\cdot)/\tau(\cdot)$. From this we see that the introduction of $\tau(\cdot)$ rescales the solution in each location, and that the variance is scaled by $\tau(\cdot)$:

$$\mathrm{Var}(u(\boldsymbol{s})) = \frac{1}{\tau^2(\boldsymbol{s})}\mathrm{Var}(\hat{u}(\boldsymbol{s})), \quad \boldsymbol{s} \in \mathbb{R}^2.$$

Further, for any $\boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathbb{R}^2$, we have that

$$\begin{aligned}\operatorname{Corr}(\hat{u}(\boldsymbol{s}_1), \hat{u}(\boldsymbol{s}_2)) &= \frac{\operatorname{Cov}(\hat{u}(\boldsymbol{s}_1), \hat{u}(\boldsymbol{s}_2))}{\operatorname{SD}(\hat{u}(\boldsymbol{s}_1))\operatorname{SD}(\hat{u}(\boldsymbol{s}_2))} \\ &= \frac{\tau(\boldsymbol{s}_1)\tau(\boldsymbol{s}_2)\operatorname{Cov}(u(\boldsymbol{s}_1), u(\boldsymbol{s}_2))}{\tau(\boldsymbol{s}_1)\operatorname{SD}(u(\boldsymbol{s}_1))\tau(\boldsymbol{s}_2)\operatorname{SD}(u(\boldsymbol{s}_2))} \\ &= \operatorname{Corr}(u(\boldsymbol{s}_1), u(\boldsymbol{s}_2)),\end{aligned}$$

where $\operatorname{SD}(\cdot)$ is the standard deviation of its argument. This means that the correlation structure is unaffected by the choice of $\tau(\cdot)$.

3.1.4 Discretization

In order to use SPDE (3.6) in a computational framework, a discrete representation of the solution is derived. This can be done in multiple ways. In Lindgren et al. (2011) a finite element representation is used, with Gaussian weights and piecewise linear basis functions. The solution is constructed on a mesh obtained by Delaunay triangulation, and is required to have a zero normal derivative along the boundary. This is known as a Neumann boundary condition. The use of triangulation makes it possible to represent arbitrarily shaped regions in $\mathbb{R}^2$ with irregularly observed locations. The resolution can also vary throughout the region, which allows for a finer level of detail where this is needed.

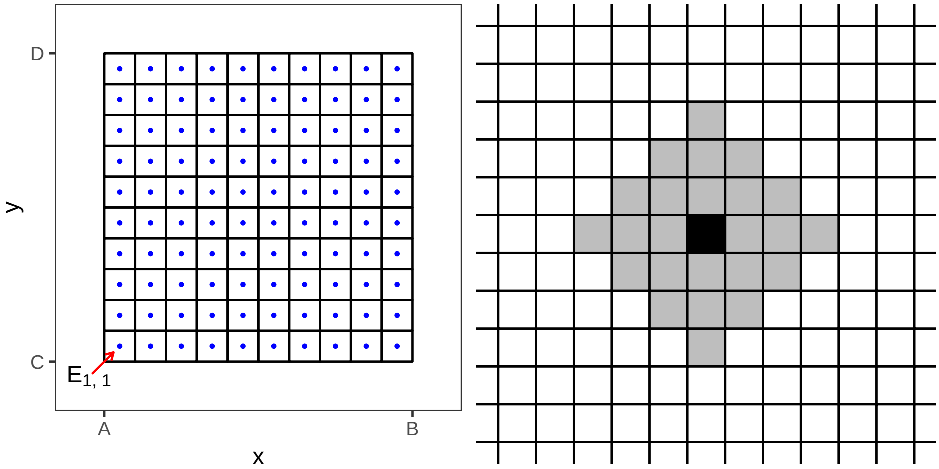
Fuglstad et al. (2015a) propose a discretization based on finite volume methods. For practial reasons, the area of interest is required to be rectangular, and the solution is approximated on a regular grid of rectangular cells. Along the boundary, periodic conditions are used. We used this representation in the project work described in Isaksen (2019), and we continue using it in this thesis. While the representation was described in the project thesis, we provide a brief summary. For a more technical derivation of the approximation, see Fuglstad et al. (2015a).

The SPDE of interest is

$$\left(\kappa^2(\boldsymbol{s}) - \nabla \cdot \mathbf{H}(\boldsymbol{s})\nabla\right)(\tau(\boldsymbol{s})u(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D},$$

where $\mathcal{D} = [A, B] \times [C, D]$ is a rectangular subdomain of $\mathbb{R}^2$ for $B > A$ and $D > C$, and the solution $u(\cdot)$ is assumed to be periodic along the vertical and horizontal boundaries. We divide $\mathcal{D}$ into a regular grid of rectangular cells with n_x and n_y cells in the x- and y-direction, respectively, resulting in a grid size of $n = n_x n_y$. The width and height of each cell is $h_x = (B - A)/n_x$ and $h_y = (D - C)/n_y$, respectively.

Define $E_{1,1}$ to be the lower left grid cell, so that $E_{i,j}$ is the grid cell in row i and column j. The center of $E_{i,j}$ is called $\boldsymbol{s}_{i,j}$. Figure 3.1a shows an example of



(a) (b)

Figure 3.1: (a) A regular 10×10 grid on the rectangle $[A, B] \times [C, D]$, with centroids shown as blue points. (b) Demonstration of the conditional independence properties of $\boldsymbol{u}$. The value in the black grid cell is conditionally independent of the white grid cells, given the grey grid cells.

a grid, where $n_x = n_y = 10$ and the centroids $\boldsymbol{s}_{i,j}$ are shown as blue points. We denote by $u_{i,j}$ the approximation to $u(\boldsymbol{s}_{i,j})$, which is equal to the average value of $u(\cdot)$ over $E_{i,j}$. By starting at $u_{1,1}$ and stacking the values of $u_{i,j}$ row-wise into the a vector, i.e., $\boldsymbol{u} = (u_{1,1}, \dots, u_{1,n_x}, u_{2,1}, \dots, u_{n_x,n_y})$, the resulting approximation satisfies $\mathbf{B}\boldsymbol{u} = \boldsymbol{z}$. Here, $\mathbf{B} \in \mathbb{R}^{n \times n}$ is the coefficient matrix of the system, and $\boldsymbol{z} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$. Solving for $\boldsymbol{u}$, we obtain

$$\boldsymbol{u} \sim \mathcal{N}_{n_x n_y}(\mathbf{0}, \mathbf{Q}^{-1}).$$

This makes $\boldsymbol{u}$ a multivariate Gaussian vector with expected value $\mathbf{0}$ and precision matrix $\mathbf{Q} = \mathbf{B}^\mathsf{T}\mathbf{B}$, which depends on $\tau(\cdot)$, $\kappa(\cdot)$, $\mathbf{H}(\cdot)$, and the area of the grid cells.

Due to the local nature of differential operators, $\mathbf{Q}$ has at most 25 non-zero elements per row. Using a grid of size of $n_x = n_y = 150$, only 0.1% of its elements are non-zero. This sparsity makes it useful to consider $\boldsymbol{u}$ a GMRF with respect to the graph defined by $\mathbf{Q}$. For each grid cell, the non-zero elements correspond to 25 neighbors of the cell, which is demonstrated in Figure 3.1b. If we condition the element of $\boldsymbol{u}$ in the black grid cell on the values of the neighboring elements, which are colored grey, then it becomes conditionally independent of the remaining elements in the white grid cells. This is an example of the local Markov property from Theorem 2.3.

At one point in the discretization, an estimate of the surface integral of $\mathbf{H}(\boldsymbol{s})\nabla u(\boldsymbol{s})$ over the boundaries of the grid cells is needed. The discretization scheme used needs to be consistent, in that the same estimate is obtained when the same boundary is considered from any two neighboring cells. The resulting estimate requires the value of $\mathbf{H}(\cdot)$ not only in the grid cells $\boldsymbol{s}_{i,j}$, but also in the locations $\boldsymbol{s}_{i,j} + (ah_x/2, bh_y/2)$ for all combinations of $a, b \in \{-1, 0, 1\}$. In other words, the computations involving $\mathbf{H}(\cdot)$ are done on a half-spaced grid. If $\mathbf{H}(\cdot)$ depends on spatial covariates, these need to be available on a finer scale than the original grid.

After discretizing, the effective ranges and marginal variance in Equations (3.4) and (3.5) no longer hold exactly. However, they become increasingly better approximations as n_x and n_y increase, which is demonstrated by numerical examples in Isaksen (2019). The periodic boundary condition introduces unwanted effects, but these are limited to the vicinity of the boundary. By letting $\mathcal{D}$ be larger than the area of interest we can ensure that these boundary effects have negligible influence.

For the non-stationary SPDEs, a closed-form expression for the marginal variance of the solution is not known, despite one being available for stationary SPDEs. As a result, the marginal variance can not be specified exactly for non-stationary SPDEs. In the discretized approximation, however, the precision matrix $\mathbf{Q}$ is obtained directly. Its inverse, the covariance matrix $\boldsymbol{\Sigma}$, can be expressed as $\boldsymbol{\Sigma} = \mathbf{D}_\sigma \boldsymbol{\Sigma}_\rho \mathbf{D}_\sigma$, where $\mathbf{D}_\sigma = \mathrm{Diag}(\boldsymbol{\Sigma})^{1/2}$ has the marginal standard deviations of $\boldsymbol{u}$ along its diagonal, and $\boldsymbol{\Sigma}_\rho$ is the correlation matrix of $\boldsymbol{u}$. In other words, we can decompose the covariance structure into the marginal variance and correlation structure. By inverting $\boldsymbol{\Sigma}$, we obtain $\mathbf{Q} = \mathbf{D}_\sigma^{-1} \boldsymbol{\Sigma}_\rho^{-1} \mathbf{D}_\sigma^{-1}$. If $\mathbf{D}_\beta$ is a diagonal matrix containing the desired marginal deviations, and $\mathbf{D} = \mathbf{D}_\beta^{-1}\mathbf{D}_\sigma$, then $\mathbf{Q}_\beta = \mathbf{DQD}$ has the wanted marginal variance:

$$\mathbf{Q}_\beta = \mathbf{DQD} = (\mathbf{D}_\beta^{-1}\mathbf{D}_\sigma)(\mathbf{D}_\sigma^{-1}\boldsymbol{\Sigma}_\rho^{-1}\mathbf{D}_\sigma^{-1})(\mathbf{D}_\beta^{-1}\mathbf{D}_\sigma) = \mathbf{D}_\beta^{-1}\boldsymbol{\Sigma}_\rho^{-1}\mathbf{D}_\beta^{-1}.$$

In this way, we can control the marginal variance without computing $\boldsymbol{\Sigma}$. The original marginal deviations in $\mathbf{D}_\sigma$ can be extracted from $\mathbf{Q}$ by an algorithm described in Section 12.1 of Gelfand et al. (2010). In the `R-INLA` library, this algorithm is implemented in the function `inla.qinv`.

3.1.5 Parametrization

There are many ways to parametrize SPDE (3.6) and variations of it. Lindgren et al. (2011) model $\log(\kappa^2(\cdot))$ and $\log(\tau(\cdot))$ as linear combinations of smooth basis functions. In Fuglstad et al. (2015a), the focus in on the effect of $\mathbf{H}(\cdot)$, with constant $\kappa(\cdot)$ and $\tau(\cdot)$ fixed to 1. Using the absolute anisotropy from Section 3.1.2, they let $\gamma(\cdot)$ be constant and model each component of $\boldsymbol{v}(\cdot)$ as a Fourier series.

In this thesis, we aim to model the non-stationarity through linear regression on spatial covariates, and, therefore, describe parametrizations that allow for this.

SPDE (3.6) depends on three functions: the scalar-valued functions $\kappa(\cdot)$ and $\tau(\cdot)$, and the matrix-valued function $\mathbf{H}(\cdot)$. For $\mathbf{H}(\cdot)$, we use the relative decomposition described in Section 3.1.2,

$$\mathbf{H}(\boldsymbol{s}) = \gamma(\boldsymbol{s})\left(\mathbf{I}_2 + \boldsymbol{w}(\boldsymbol{s})\boldsymbol{w}(\boldsymbol{s})^\mathsf{T}\right), \quad \boldsymbol{s} \in \mathcal{D}$$

While we know, qualitatively, how $\kappa(\cdot), \tau(\cdot)$, and $\gamma(\cdot)$ affect the solution, the functions in themselves do not represent intuitive quantities. For the sake of comprehensibility, we let the regressions be on functions that have a clear interpretation. For example, in Parametrization S-NS1 we let the approximate baseline effective range function $\rho(\cdot) = \sqrt{8\gamma(\cdot)}$ be modeled by regression, instead of $\gamma(\cdot)$. This also makes it easier to specify sensible priors.

Below, five parametrizations are described. Note that, in order to obtain an identifiable model, either $\kappa(\cdot)$ or $\tau(\cdot)$ is fixed to 1 in each parametrization. S-ISO and S-ANISO lead to stationary models, with the former isotropic and the latter geometrically anisotropic. S-NS1, S-NS1E, and S-NS2 are non-stationary models, and differ mainly in the way they control the marginal variance. In S-NS1 and S-NS1E we fix $\kappa(\cdot)$. The correlation structure is first specified by $\mathbf{H}(\cdot)$, and then $\tau(\cdot)$ is chosen so that it corrects for the spatially varying marginal variance introduced by $\mathbf{H}(\cdot)$. In S-NS2 we fix $\tau(\cdot)$, and specify $\mathbf{H}(\cdot)$ and $\kappa(\cdot)$ simultaneously, so that both control the correlation structure and marginal variance. In the descriptions of the parametrizations, we assume that both the region $\mathcal{D}$ and its rectangular grid with centroids $\{\boldsymbol{s}_i : i = 1, \ldots, n_x n_y\}$ have been specified.

Parametrization S-ISO. We fix $\kappa(\boldsymbol{s}) \equiv 1$, $\mathbf{H}(\boldsymbol{s}) \equiv \mathbf{H}$ and $\tau(\boldsymbol{s}) \equiv \tau$ for $\boldsymbol{s} \in \mathcal{D}$, and let

$$\mathbf{H} = \frac{1}{8}\rho^2\mathbf{I}_2,$$

where ρ is the effective range. The marginal variance of the solution is then $\sigma^2 = 2/(\pi\tau^2\rho^2)$. We obtain this marginal variance by letting

$$\tau = \sqrt{\frac{2}{\pi}}\frac{1}{\sigma\rho}.$$

Since the discretized GMRF is an approximation to the solution of the SPDE, the parameter σ^2 is not equal to the exact marginal variance of the GMRF. We control the parametrization through two parameters, namely $\boldsymbol{\theta} = (\rho, \sigma)$. Since both of these are required to be positive, we use $\log(\rho)$ and $\log(\sigma)$ during computations. △

Parametrization S-ANISO. We fix $\kappa(\boldsymbol{s}) \equiv 1$, $\mathbf{H}(\boldsymbol{s}) \equiv \mathbf{H}$ and $\tau(\boldsymbol{s}) \equiv \tau$ for $\boldsymbol{s} \in \mathcal{D}$, and let

$$\mathbf{H} = \frac{1}{8}\rho^2 \left(\mathbf{I}_2 + \boldsymbol{w}\boldsymbol{w}^\mathsf{T}\right),$$

where ρ is the baseline effective range and $\boldsymbol{w} = (w_x, w_y)$ controls the direction of longest range, and the strength of the anisotropy. The marginal variance is then $\sigma^2 = 2/\left(\pi\tau^2\rho^2\sqrt{1 + w_x^2 + w_y^2}\right)$, which is obtained by letting

$$\tau = \sqrt{\frac{2}{\pi}}\frac{1}{\sigma\rho}\left[1 + w_x^2 + w_y^2\right]^{-1/4}.$$

Since the discretized GMRF is an approximation to the solution of the SPDE, the parameter σ^2 is not equal to the exact marginal variance of the GMRF. We control the parametrization through four parameters, namely $\boldsymbol{\theta} = (\rho, \sigma, w_x, w_y)$. Since ρ and σ are required to be positive, we use $\log(\rho)$ and $\log(\sigma)$ during computations. △

Parametrization S-NS1 & S-NS1E. We fix $\kappa(\boldsymbol{s}) \equiv 1$ for $\boldsymbol{s} \in \mathcal{D}$, but let $\mathbf{H}(\cdot)$ and $\tau(\cdot)$ be spatially varying functions. The matrix-valued function $\mathbf{H}(\cdot)$ is parametrized as

$$\mathbf{H}(\boldsymbol{s}) = \frac{1}{8}\rho^2(\boldsymbol{s})\left(\mathbf{I}_2 + \boldsymbol{w}(\boldsymbol{s})\boldsymbol{w}(\boldsymbol{s})^\mathsf{T}\right), \quad \boldsymbol{s} \in \mathcal{D},$$

where $\rho(\cdot)$ is the baseline effective range function and $\boldsymbol{w}(\cdot) = (w_x(\cdot), w_y(\cdot))$ controls the direction and the magnitude of the anisotropy throughout the region. In order to specify $\tau(\cdot)$, we use the intuition from Section 3.1.3 and consider the SPDE

$$(1 - \nabla \cdot \mathbf{H}(\boldsymbol{s})\nabla)\,\hat{u}(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D}. \tag{3.10}$$

The solution $\hat{u}(\cdot)$ has the desired correlation structure, but a varying marginal variance that depends on $\mathbf{H}(\cdot)$, given by $\sigma_{\mathbf{H}}^2(\boldsymbol{s}) = \mathrm{Var}(\hat{u}(\boldsymbol{s}))$ for $\boldsymbol{s} \in \mathcal{D}$. If we let $\tau(\boldsymbol{s}) = \sigma_{\mathbf{H}}(\boldsymbol{s})/\sigma(\boldsymbol{s})$, where $\sigma^2(\cdot)$ is the desired marginal variance, then $u(\boldsymbol{s}) = \hat{u}(\boldsymbol{s})/\tau(\boldsymbol{s})$ satisfies $\mathrm{Var}(u(\boldsymbol{s})) = \sigma^2(\boldsymbol{s})$ for $\boldsymbol{s} \in \mathcal{D}$.

The functions $\rho(\cdot), w_x(\cdot), w_y(\cdot)$, and $\sigma(\cdot)$ are modeled by linear regression on spatial covariates, leading to

$$\begin{aligned}
\log(\rho(\boldsymbol{s})) &= \log(\rho_0) + \boldsymbol{x}_\rho(\boldsymbol{s})^\mathsf{T}\boldsymbol{\beta}_\rho,\\
w_x(\boldsymbol{s}) &= w_{x,0} + \boldsymbol{x}_{w_x}(\boldsymbol{s})^\mathsf{T}\boldsymbol{\beta}_{w_x},\\
w_y(\boldsymbol{s}) &= w_{y,0} + \boldsymbol{x}_{w_y}(\boldsymbol{s})^\mathsf{T}\boldsymbol{\beta}_{w_y},\\
\log(\sigma(\boldsymbol{s})) &= \log(\sigma_0) + \boldsymbol{x}_\sigma(\boldsymbol{s})^\mathsf{T}\boldsymbol{\beta}_\sigma,
\end{aligned}$$

for $\boldsymbol{s} \in \mathcal{D}$, where $\rho(\cdot)$ and $\sigma(\cdot)$ are modeled on the log-level to ensure positivity. For each function $\phi(\cdot)$, $\boldsymbol{x}_\phi(\boldsymbol{s}) \in \mathbb{R}^{p_\phi}$ is a vector containing the spatial covariates at location $\boldsymbol{s} \in \mathcal{D}$, and $\boldsymbol{\beta}_\phi \in \mathbb{R}^{p_\phi}$ quantifies the effect of each covariate. The model parameters are $\boldsymbol{\theta} = (\rho_0, w_{x_0}, w_{y,0}, \sigma_0, \boldsymbol{\beta}_\rho^\mathsf{T}, \boldsymbol{\beta}_{w_x}^\mathsf{T}, \boldsymbol{\beta}_{w_y}^\mathsf{T}, \boldsymbol{\beta}_\sigma^\mathsf{T})$, leading to $4 + p_\rho + p_{w_x} + p_{w_y} + p_\sigma$ parameters in total. Here p_ρ, p_{w_x}, p_{w_y}, and p_σ are the number of spatial covariates used in each corresponding function.

This parametrization relies on knowing $\sigma^2_\mathbf{H}(\cdot)$, the marginal variance function of the solution to SPDE (3.10). In general, we do not have a closed-form expression for this. We propose two ways of controlling the marginal variance:

$\boxed{\text{S-NS1E}}$ Construct the precision matrix $\mathbf{Q}_\mathbf{H}$ of the GMRF approximation $\hat{\boldsymbol{u}}$ to $\hat{u}(\cdot)$, and compute the partial inverse $\boldsymbol{\Sigma}^*$. The diagonal elements of $\boldsymbol{\Sigma}^*$ are the marginal variances of $\hat{\boldsymbol{u}}$, which are then used as values of $\sigma^2_\mathbf{H}(\cdot)$ on the grid for $\mathcal{D}$. The computation of $\boldsymbol{\Sigma}^*$ has a computational complexity of $\mathcal{O}((n_x n_y)^{1.5})$.

$\boxed{\text{S-NS1}}$ Let $\sigma^2_\mathbf{H}(\cdot)$ be the approximate marginal variance of $\hat{u}(\cdot)$ from Equation (3.8), i.e.,

$$\sigma^2_\mathbf{H}(\boldsymbol{s}) = \frac{1}{4\pi\kappa^2(\boldsymbol{s})\tau^2(\boldsymbol{s})\left|\mathbf{H}(\boldsymbol{s})\right|^{1/2}} = \frac{2}{\pi\rho^2(\boldsymbol{s})\sqrt{1 + w_x^2(\boldsymbol{s}) + w_y^2(\boldsymbol{s})}}, \quad \boldsymbol{s} \in \mathcal{D}.$$

The closed form for $\tau(\cdot)$ is then

$$\tau(\boldsymbol{s}) = \sqrt{\frac{2}{\pi}}\frac{1}{\sigma(\boldsymbol{s})\rho(\boldsymbol{s})}\left[1 + w_x^2(\boldsymbol{s}) + w_y^2(\boldsymbol{s})\right]^{-1/4}, \quad \boldsymbol{s} \in \mathcal{D}.$$

S-NS1E leads to exact control of the marginal variance, while S-NS1 gives approximate control. However, S-NS1E requires the partial inverse $\boldsymbol{\Sigma}^*$, which is expensive to compute. The approximation used in S-NS1 can be computed in $\mathcal{O}(n_x n_y)$ time. △

Parametrization S-NS2. We fix $\tau(\boldsymbol{s}) \equiv 1$ for $\boldsymbol{s} \in \mathcal{D}$, but let $\mathbf{H}(\cdot)$ and $\kappa(\cdot)$ be spatially varying functions. $\mathbf{H}(\cdot)$ is parametrized as

$$\mathbf{H}(\boldsymbol{s}) = \gamma(\boldsymbol{s})\left(\mathbf{I}_2 + \boldsymbol{w}(\boldsymbol{s})\boldsymbol{w}(\boldsymbol{s})^\mathsf{T}\right), \quad \boldsymbol{s} \in \mathcal{D}.$$

Further, we define the functions $\rho(\cdot)$ and $\sigma^2(\cdot)$ to be given by

$$\rho(\boldsymbol{s}) = \frac{\sqrt{8}}{\kappa(\boldsymbol{s})}\sqrt{\gamma(\boldsymbol{s})} \text{ and } \sigma^2(\boldsymbol{s}) = \frac{1}{4\pi\kappa^2(\boldsymbol{s})\gamma(\boldsymbol{s})\sqrt{1 + w_x^2(\boldsymbol{s}) + w_y^2(\boldsymbol{s})}}.$$

$\rho(\cdot)$ is the approximate baseline effective range from Equation (3.7), and $\sigma^2(\cdot)$ is the approximate marginal variance from Equation (3.8). Based on these two expressions, we can solve for $\kappa^2(\cdot)$ and $\gamma(\cdot)$, leading to

$$\begin{aligned}\kappa^2(\boldsymbol{s}) &= \sqrt{\frac{2}{\pi}}\frac{1}{\rho(\boldsymbol{s})\sigma(\boldsymbol{s})}\left[1+w_x^2(\boldsymbol{s})+w_y^2(\boldsymbol{s})\right]^{-1/4}, \quad \boldsymbol{s}\in\mathcal{D},\\ \gamma(\boldsymbol{s}) &= \frac{1}{\sqrt{32\pi}}\frac{\rho(\boldsymbol{s})}{\sigma(\boldsymbol{s})}\left[1+w_x^2(\boldsymbol{s})+w_y^2(\boldsymbol{s})\right]^{-1/4}, \quad \boldsymbol{s}\in\mathcal{D}.\end{aligned}$$

The functions $\rho(\cdot), w_x(\cdot), w_y(\cdot)$, and $\sigma(\cdot)$ are modeled by linear regression on spatial covariates, leading to

$$\begin{aligned}\log(\rho(\boldsymbol{s})) &= \log(\rho_0) + \boldsymbol{x}_\rho(\boldsymbol{s})^\mathsf{T}\boldsymbol{\beta}_\rho,\\ w_x(\boldsymbol{s}) &= w_{x,0} + \boldsymbol{x}_{w_x}(\boldsymbol{s})^\mathsf{T}\boldsymbol{\beta}_{w_x},\\ w_y(\boldsymbol{s}) &= w_{y,0} + \boldsymbol{x}_{w_y}(\boldsymbol{s})^\mathsf{T}\boldsymbol{\beta}_{w_y},\\ \log(\sigma(\boldsymbol{s})) &= \log(\sigma_0) + \boldsymbol{x}_\sigma(\boldsymbol{s})^\mathsf{T}\boldsymbol{\beta}_\sigma,\end{aligned}$$

for $\boldsymbol{s}\in\mathcal{D}$, where $\rho(\cdot)$ and $\sigma(\cdot)$ are modeled on the log-level to ensure positivity. For each function $\phi(\cdot)$, $\boldsymbol{x}_\phi(\boldsymbol{s})\in\mathbb{R}^{p_\phi}$ is a vector containing the spatial covariates at location $\boldsymbol{s}\in\mathcal{D}$, and $\boldsymbol{\beta}_\phi\in\mathbb{R}^{p_\phi}$ quantifies the effect of each covariate. The parameters needed for this parametrization are $\boldsymbol{\theta} = (\rho_0, w_{x_0}, w_{y,0}, \sigma_0, \boldsymbol{\beta}_\rho^\mathsf{T}, \boldsymbol{\beta}_{w_x}^\mathsf{T}, \boldsymbol{\beta}_{w_y}^\mathsf{T}, \boldsymbol{\beta}_\sigma^\mathsf{T})$, resulting in $4 + p_\rho + p_{w_x} + p_{w_y} + p_\sigma$ parameters in total. Here p_ρ, p_{w_x}, p_{w_y}, and p_σ are the number of spatial covariates used in each corresponding function. △

3.2 Kernel convolutions

3.2.1 Non-stationary covariance functions

The following approach is based on the work first described in Higdon et al. (1999), where the focus is on GRFs that can be expressed as the convolution between a kernel and a noise process:

$$u(\boldsymbol{s}) = \int_{\mathbb{R}^2} K_{\boldsymbol{s}}(\boldsymbol{x})\mathcal{W}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \quad \boldsymbol{s}\in\mathbb{R}^2.$$

The kernel $K_{\boldsymbol{s}}(\cdot)$ is centered at $\boldsymbol{s}$, and has a shape that can vary as a function of $\boldsymbol{s}$, and $\mathcal{W}(\cdot)$ is a standard Gaussian white noise process. For $\boldsymbol{s}_1, \boldsymbol{s}_2\in\mathbb{R}^2$, the covariance is given by

$$C(\boldsymbol{s}_1,\boldsymbol{s}_2) = \mathrm{Cov}(u(\boldsymbol{s}_1), u(\boldsymbol{s}_2)) = \int_{\mathbb{R}^2} K_{\boldsymbol{s}_1}(\boldsymbol{x})K_{\boldsymbol{s}_2}(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

and consequently depends on the amount of overlap between the kernels at each pair of locations. This covariance function is valid regardless of how the kernel $K_{\boldsymbol{s}}(\cdot)$ is chosen, as long as we ensure that the marginal variance is bounded, i.e., that $\sup_{\boldsymbol{s}\in\mathbb{R}^2}\int_{\mathbb{R}^2} K_{\boldsymbol{s}}^2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} < \infty$ (Higdon et al., 1999).

Instead of specifying the covariance function directly, we control it indirectly through the kernel function. This leads to flexible control of the covariance structure, without having to worry about positive definiteness. While the choice of kernel function is arbitrary, Gaussian kernels on the form

$$K_{\boldsymbol{s}}(\boldsymbol{x}) = \frac{1}{2\pi}\sigma(\boldsymbol{s})\left|\boldsymbol{\Sigma}(\boldsymbol{s})\right|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{s})^{\mathsf{T}}\boldsymbol{\Sigma}(\boldsymbol{s})^{-1}(\boldsymbol{x}-\boldsymbol{s})\right], \quad \boldsymbol{s}\in\mathbb{R}^2$$

are a popular alternative. Here, $\sigma(\cdot)$ is the marginal standard deviation function and $\boldsymbol{\Sigma}(\cdot)$ is the *kernel matrix* function. For all $\boldsymbol{s}\in\mathbb{R}^2$, $\boldsymbol{\Sigma}(\boldsymbol{s})$ is a 2×2 positive definite matrix. This leads to a kernel that decays monotonically along all directions from the mode at $\boldsymbol{s}$, and whose shape is determined by the matrix $\boldsymbol{\Sigma}(\boldsymbol{s})$.

As shown in Paciorek and Schervish (2006), this kernel choice leads to a covariance function with a closed form given by

$$C(\boldsymbol{s}_1,\boldsymbol{s}_2) = \sigma(\boldsymbol{s}_1)\sigma(\boldsymbol{s}_2)\frac{\left|\boldsymbol{\Sigma}(\boldsymbol{s}_1)\right|^{1/4}\left|\boldsymbol{\Sigma}(\boldsymbol{s}_2)\right|^{1/4}}{\left|\frac{\boldsymbol{\Sigma}(\boldsymbol{s}_1)+\boldsymbol{\Sigma}(\boldsymbol{s}_2)}{2}\right|^{1/2}}\exp\left(-Q(\boldsymbol{s}_1,\boldsymbol{s}_2)\right) \tag{3.11}$$

for $\boldsymbol{s}_1,\boldsymbol{s}_2\in\mathbb{R}^2$, where

$$Q(\boldsymbol{s}_1,\boldsymbol{s}_2) = (\boldsymbol{s}_1-\boldsymbol{s}_2)^{\mathsf{T}}\left(\frac{\boldsymbol{\Sigma}(\boldsymbol{s}_1)+\boldsymbol{\Sigma}(\boldsymbol{s}_2)}{2}\right)^{-1}(\boldsymbol{s}_1-\boldsymbol{s}_2)$$

is the Mahalanobis distance between $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ with respect to $(\boldsymbol{\Sigma}(\boldsymbol{s}_1)+\boldsymbol{\Sigma}(\boldsymbol{s}_2))/2$. In other words, the covariance between any two locations is fully determined by their relative positions, and the value of $\sigma(\cdot)$ and $\boldsymbol{\Sigma}(\cdot)$ in each location.

As long as the kernel $K_{\boldsymbol{s}}(\cdot)$ varies smoothly as a function of $\boldsymbol{s}$, realizations from GRFs with this covariance function are infinitely differentiable. The issues associated with covariance functions leading to such smooth realizations are discussed in Chapter 3 of Stein (1999). In order to bypass this limitation, a modified version of Equation (3.11) is considered. If $\rho(\cdot)$ is an isotropic correlation function, i.e., a positive definite function with $\rho(0)=1$, then

$$C(\boldsymbol{s}_1,\boldsymbol{s}_2) = \sigma(\boldsymbol{s}_1)\sigma(\boldsymbol{s}_2)\frac{\left|\boldsymbol{\Sigma}(\boldsymbol{s}_1)\right|^{1/4}\left|\boldsymbol{\Sigma}(\boldsymbol{s}_2)\right|^{1/4}}{\left|\frac{\boldsymbol{\Sigma}(\boldsymbol{s}_1)+\boldsymbol{\Sigma}(\boldsymbol{s}_2)}{2}\right|^{1/2}}\rho\left(\sqrt{Q(\boldsymbol{s}_1,\boldsymbol{s}_2)}\right), \quad \boldsymbol{s}_1,\boldsymbol{s}_2\in\mathbb{R}^2,$$

is a valid covariance function (Paciorek and Schervish, 2006). In Paciorek and Schervish (2004), the Matérn correlation function is used for $\rho(\cdot)$, leading to the covariance function

$$C(\boldsymbol{s}_1, \boldsymbol{s}_2) = \sigma(\boldsymbol{s}_1)\sigma(\boldsymbol{s}_2)\frac{|\boldsymbol{\Sigma}(\boldsymbol{s}_1)|^{1/4}\,|\boldsymbol{\Sigma}(\boldsymbol{s}_2)|^{1/4}}{\left|\frac{\boldsymbol{\Sigma}(\boldsymbol{s}_1)+\boldsymbol{\Sigma}(\boldsymbol{s}_2)}{2}\right|^{1/2}}\mathcal{M}_\nu\left(\sqrt{Q(\boldsymbol{s}_1, \boldsymbol{s}_2)}\right), \tag{3.12}$$

for $\boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathbb{R}^2$. The parametrization used for the Matérn correlation function is

$$\mathcal{M}_\nu(h) = \frac{1}{\Gamma(\nu)2^{\nu-1}}h^\nu K_\nu(h), \quad h \geq 0,$$

so that the range and marginal standard deviation are determined by $\boldsymbol{\Sigma}(\cdot)$ and $\sigma(\cdot)$, respectively. More flexibility can be attained by letting ν vary spatially as well, but we restrict ourselves to the constant case. While the value of ν can be estimated from observed data, we treat it as a known constant and fix its value prior to inference. Since the SPDE-based approach in Section 3.1 leads to a Matérn GRF with $\nu = 1$, we will use this value for the kernel-based approach as well.

When both $\sigma(\cdot)$ and $\boldsymbol{\Sigma}(\cdot)$ are constant, with $\sigma(\boldsymbol{s}) \equiv \sigma$ and $\boldsymbol{\Sigma}(\boldsymbol{s}) \equiv \boldsymbol{\Sigma}$ for all $\mathbf{s} \in \mathbb{R}^2$, the covariance function is given by

$$C(\boldsymbol{s}_1, \boldsymbol{s}_2) = \sigma^2\mathcal{M}_1\left(\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{s}_1 - \boldsymbol{s}_2)\|\right).$$

This is a stationary covariance function. By comparison with Equation (3.3), we see that $\boldsymbol{\Sigma}$ then plays the same role as $\mathbf{H}$ in SPDE (3.2) - it determines the direction and strength of the anisotropy. By letting $\boldsymbol{\Sigma}(\cdot)$ vary spatially, we obtain a GRF with spatially varying anisotropy properties. Analogous to the stationary case, $\boldsymbol{\Sigma}(\cdot)$ controls the properties of the correlation structure throughout $\mathbb{R}^2$, such as the baseline range, the strength of the anisotropy, and the direction of longest range.

3.2.2 Parametrization

The covariance function in Equation (3.12) is specified by two functions: $\sigma(\cdot)$ and $\boldsymbol{\Sigma}(\cdot)$. Several approaches for modeling these functions have been proposed. In Higdon et al. (1999), $\sigma(\cdot)$ is a spatial constant and $\boldsymbol{\Sigma}(\boldsymbol{s})$ for $\boldsymbol{s} \in \mathbb{R}^2$ is controlled indirectly by considering the level curves of the resulting kernel $K_{\boldsymbol{s}}(\cdot)$, which are ellipses. For this purpose, they focus on the *one standard deviation ellipse* $E_{\boldsymbol{s}}$, which satisfies

$$\frac{\int_{A_{\boldsymbol{s}}} K_{\boldsymbol{s}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}{\int_{\mathbb{R}^2} K_{\boldsymbol{s}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}} \approx 0.68,$$

where $A_{\boldsymbol{s}}$ is the region enclosed by $E_{\boldsymbol{s}}$. The kernel is centered at $\boldsymbol{s}$, so that the foci of $E_{\boldsymbol{s}}$ are $\boldsymbol{s} \pm \boldsymbol{\psi}$, where $\boldsymbol{\psi} = (\psi_x, \psi_y)$. The components of $\boldsymbol{\psi}$ are modeled as independent stationary GRFs, and a unique ellipse is obtained by fixing $|A_{\boldsymbol{s}}| = A$ for all $\boldsymbol{s} \in \mathbb{R}^2$, where A is chosen prior to modeling. Paciorek and Schervish (2006) let $\sigma(\cdot)$ be a spatially constant function and use the eigendecomposition of $\boldsymbol{\Sigma}(\boldsymbol{s})$,

$$\boldsymbol{\Sigma}(\mathrm{s}) = \boldsymbol{\Gamma}(\boldsymbol{s})\boldsymbol{\Lambda}(\boldsymbol{s})\boldsymbol{\Gamma}(\boldsymbol{s})^{\mathsf{T}},$$

where

$$\boldsymbol{\Lambda}(\boldsymbol{s}) = \begin{pmatrix} \lambda_1(\boldsymbol{s}) & 0 \\ 0 & \lambda_2(\boldsymbol{s}) \end{pmatrix}, \quad \boldsymbol{\Gamma}(\boldsymbol{s}) = \begin{pmatrix} \cos\xi(\boldsymbol{s}) & -\sin\xi(\boldsymbol{s}) \\ \sin\xi(\boldsymbol{s}) & \cos\xi(\boldsymbol{s}) \end{pmatrix},$$

and model $\lambda_1(\cdot), \lambda_2(\cdot)$, and $\xi(\cdot)$ using an approximate stationary GRF representation.

Hoff and Niu (2012) propose a covariance regression model, which, in our context, can be formulated as

$$\boldsymbol{\Sigma}(\boldsymbol{s}) = \boldsymbol{\Psi} + \boldsymbol{\Gamma}\boldsymbol{x}_{\boldsymbol{\Sigma}}(\boldsymbol{s})\boldsymbol{x}_{\boldsymbol{\Sigma}}(\boldsymbol{s})^{\mathsf{T}}\boldsymbol{\Gamma}^{\mathsf{T}}.$$

Here, $\boldsymbol{\Psi} \in \mathbb{R}^{2\times 2}$ is a positive definite matrix specifying the "baseline" anisotropy, $\boldsymbol{x}_{\boldsymbol{\Sigma}}(\boldsymbol{s}) \in \mathbb{R}^{p_{\boldsymbol{\Sigma}}}$ is a vector of spatial covariates at $\boldsymbol{s}$ and $\boldsymbol{\Gamma} \in \mathbb{R}^{2\times p_{\boldsymbol{\Sigma}}}$ is matrix of regression coefficients, determining the effect of the covariates on $\boldsymbol{\Sigma}(\cdot)$.

In order to get a fair basis of comparison between the SPDE- and kernel-based approaches, we should ensure that the parametrizations used are as similar as possible. Since $\boldsymbol{\Sigma}(\cdot)$ plays a similar role as $\mathbf{H}(\cdot)$ in SPDE (3.6), we propose a decomposition analogous to the one described in Section 3.1.2:

$$\boldsymbol{\Sigma}(\boldsymbol{s}) = \gamma(\boldsymbol{s})\left(\mathbf{I}_2 + \boldsymbol{w}(\boldsymbol{s})\boldsymbol{w}(\boldsymbol{s})^{\mathsf{T}}\right), \quad \boldsymbol{s} \in \mathbb{R}^2, \tag{3.13}$$

where $\gamma(\cdot)$ is a strictly positive function and $\boldsymbol{w}(\cdot) = (w_x(\cdot), w_y(\cdot))$.

Below, three parametrizations are described: the stationary K-ISO and K-ANISO, and the non-stationary K-NS. K-ISO and K-ANISO lead to isotropic and geometrically anisotropic covariance structures, respectively. These parametrizations lead to the same models as S-ISO and S-ANISO, ignoring the approximation differences between the SPDE- and kernel-based approaches. K-NS models $\sigma(\cdot)$ and the components of $\boldsymbol{\Sigma}(\cdot)$ through regression on spatial covariates, and corresponds to S-NS1, S-NS1E and S-NS2.

Parametrization K-ISO. We fix $\sigma(\boldsymbol{s}) \equiv \sigma$ and $\boldsymbol{\Sigma}(\boldsymbol{s}) \equiv \boldsymbol{\Sigma}$ for all $\boldsymbol{s} \in \mathbb{R}$. $\boldsymbol{\Sigma}$ is parametrized as

$$\boldsymbol{\Sigma} = \frac{1}{8}\rho^2\mathbf{I}_2,$$

where ρ is the effective range. The parametrization contains two parameters, namely $\boldsymbol{\theta} = (\rho, \sigma)$. Computations are done with $\log(\rho)$ and $\log(\sigma)$ to ensure positivity. △

Parametrization K-ANISO. We fix $\sigma(\boldsymbol{s}) \equiv \sigma$ and $\boldsymbol{\Sigma}(\boldsymbol{s}) \equiv \boldsymbol{\Sigma}$ for all $\boldsymbol{s} \in \mathbb{R}$. $\boldsymbol{\Sigma}$ is parametrized as

$$\boldsymbol{\Sigma} = \frac{1}{8}\rho^2 \left(\mathbf{I}_2 + \boldsymbol{w}\boldsymbol{w}^\mathsf{T}\right),$$

where ρ is the baseline effective range and $\boldsymbol{w} = (w_x, w_y)$ specifies the direction and magnitude of the anisotropy. The parametrization contains four parameters, namely $\boldsymbol{\theta} = (\rho, \sigma, w_x, w_y)$. Computations are done with $\log(\rho)$ and $\log(\sigma)$ to ensure positivity. △

Parametrization K-NS. Both $\sigma(\cdot)$ and $\boldsymbol{\Sigma}(\cdot)$ are allowed to vary spatially. We parametrize the matrix valued function $\boldsymbol{\Sigma}(\cdot)$ as

$$\boldsymbol{\Sigma}(\boldsymbol{s}) = \frac{1}{8}\rho^2(\boldsymbol{s}) \left(\mathbf{I}_2 + \boldsymbol{w}(\boldsymbol{s})\boldsymbol{w}(\boldsymbol{s})^\mathsf{T}\right), \quad \boldsymbol{s} \in \mathbb{R}^2,$$

where $\rho(\cdot)$ is the approximate baseline effective range function, and $\boldsymbol{w}(\cdot) = (w_x(\cdot), w_y(\cdot))$ controls the direction and magnitude of the anisotropy throughout $\mathbb{R}^2$. The functions $\rho(\cdot), w_x(\cdot), w_y(\cdot)$, and $\sigma(\cdot)$ are modeled by regression on spatial covariates, leading to

$$\begin{aligned}
\log(\rho(\boldsymbol{s})) &= \log(\rho_0) + \boldsymbol{x}_\rho(\boldsymbol{s})^\mathsf{T}\boldsymbol{\beta}_\rho, \\
w_x(\boldsymbol{s}) &= w_{x,0} + \boldsymbol{x}_{w_x}(\boldsymbol{s})^\mathsf{T}\boldsymbol{\beta}_{w_x}, \\
w_y(\boldsymbol{s}) &= w_{y,0} + \boldsymbol{x}_{w_y}(\boldsymbol{s})^\mathsf{T}\boldsymbol{\beta}_{w_y}, \\
\log(\sigma(\boldsymbol{s})) &= \log(\sigma_0) + \boldsymbol{x}_\sigma(\boldsymbol{s})^\mathsf{T}\boldsymbol{\beta}_\sigma,
\end{aligned}$$

for $\boldsymbol{s} \in \mathbb{R}^2$. We ensure that $\rho(\cdot)$ and $\sigma(\cdot)$ are positive by letting the regression be on the log-level. For each function $\phi(\cdot)$, $\boldsymbol{x}_\phi(\boldsymbol{s}) \in \mathbb{R}^{p_\phi}$ is a vector containing the spatial covariates at location $\boldsymbol{s} \in \mathcal{D}$, and $\boldsymbol{\beta}_\phi \in \mathbb{R}^{p_\phi}$ quantifies the effect of each covariate. The parameters needed for this parametrization are $\boldsymbol{\theta} = (\rho_0, w_{x_0}, w_{y,0}, \sigma_0, \boldsymbol{\beta}_\rho^\mathsf{T}, \boldsymbol{\beta}_{w_x}^\mathsf{T}, \boldsymbol{\beta}_{w_y}^\mathsf{T}, \boldsymbol{\beta}_\sigma^\mathsf{T})$, resulting in $4 + p_\rho + p_{w_x} + p_{w_y} + p_\sigma$ parameters in total. Here p_ρ, p_{w_x}, p_{w_y}, and p_σ are the number of spatial covariates used in each corresponding function. △

3.3 Prior distributions

Since we wish to use these spatial models in a Bayesian framework, it is necessary to choose prior distributions for the model parameters. A useful principle when specifying priors is *Occam's razor.* It states that, when multiple competing explanations for a phenomenon are available, the simplest one is more likely to be correct than the rest. In more technical terms: if multiple models explain the observed data comparatively well, the most parsimonious[1] model should be chosen. We therefore specify priors that penalize complex models, and reward those that are close to the baseline model[2]. In this way, the prior will have a regulative effect and ensure that the posterior estimates of superfluous parameters are shrunk towards 0. This is particularly important when the model is flexible and the size of the observed data is small.

In this context, the baseline model is isotropic. Every parametrization except S-ISO and K-ISO leads to a model that, in general, is not isotropic. In these parametrizations, any parameter that leads to deviation from isotropy is given a centered Gaussian prior, which places the prior mode of the parameter at 0. This means that, prior to observing the data, the most likely value of the parameter is 0. We describe the priors used for each parametrization below.

S-ISO and K-ISO

The parameters ρ and σ are both required to be positive, and are therefore given log-normal priors:

$$\log(\rho) \sim \mathcal{N}(\mu_\rho, v_\rho^2), \quad \log(\sigma) \sim \mathcal{N}(\mu_\sigma, v_\sigma^2).$$

As a result, e^{μ_ρ} and e^{μ_σ} specify the median of the corresponding parameter, while v_ρ and v_σ control the spread.

S-ANISO and K-ANISO

In addition to having log-normal priors for ρ and σ as above, we let the both components of $\boldsymbol{w}$ have centered Gaussian priors:

$$\boldsymbol{w} \sim \mathcal{N}_2(\mathbf{0}, v_{\boldsymbol{w}}^2 \mathbf{I}_2).$$

S-NS1, S-NS1E, S-NS2, and K-NS

The parameters ρ_0 and σ_0 in the non-stationary parametrizations correspond to ρ and σ in the stationary parametrizations. Therefore, they are equipped with log-normal priors:

$$\log(\rho_0) \sim \mathcal{N}(\mu_\rho, v_\rho^2), \quad \log(\sigma_0) \sim \mathcal{N}(\mu_\sigma, v_\sigma^2).$$

Similarly, $w_{x,0}$ and $w_{y,0}$ play the same role as w_x and w_y in S-ANISO and K-ANISO, and are therefore given the prior

$$(w_{x,0}, w_{y,0}) \sim \mathcal{N}_2(\mathbf{0}, v_{\boldsymbol{w}}^2 \mathbf{I}_2).$$

[1] A model is said to be more parsimonious than another if it contains fewer parameters.

[2] By baseline model, we mean the most parsimonious model available.

The remaining parameters are the regression coefficients of $\rho(\cdot), w_x(\cdot), w_y(\cdot)$, and $\sigma(\cdot)$. We collect these in the vector $\boldsymbol{\theta}_{\mathrm{NS}} = (\boldsymbol{\beta}_\rho^\mathsf{T}, \boldsymbol{\beta}_{w_x}^\mathsf{T}, \boldsymbol{\beta}_{w_y}^\mathsf{T}, \boldsymbol{\beta}_\sigma^\mathsf{T})$, which is then given the prior

$$\boldsymbol{\theta}_{\mathrm{NS}} \sim \mathcal{N}_{p_{\mathrm{NS}}}(\mathbf{0}, v_{\mathrm{NS}}^2 \mathbf{I}_{p_{\mathrm{NS}}}),$$

where $p_{\mathrm{NS}} = p_\rho + p_{w_x} + p_{w_y} + p_\rho$ are the total number of coefficients. In other words, we let all of the regression coefficients have Gaussian priors with the same standard deviation. Therefore, the different spatial covariates should be on a similar scale, which can be achieved by standardizing them in advance.

Hyperparameters

The priors described above are choices of families of distributions that depend on hyperparameters. Sometimes, the hyperparameters can be chosen in a sensible way by using the interpretation of the parameter. Consider, for example, the parametrizations S-ISO and K-ISO. Both the effective range ρ and marginal standard deviation σ have log-normal priors. If $X \sim \mathrm{Lognormal}(\mu_X, v_X^2)$, then we can ensure that

$$\mathrm{P}(x_{\mathrm{lower}} < X < x_{\mathrm{upper}}) = 1 - \alpha$$

by choosing[3]

$$\mu_X = \frac{\log(x_{\mathrm{upper}}) + \log(x_{\mathrm{lower}})}{2}, \quad v_X = \frac{\log(x_{\mathrm{upper}}) - \log(x_{\mathrm{lower}})}{2z_{\alpha/2}}.$$

In words, we can specify the distribution of X by first identifying a reasonable prior credible interval. For the effective range, the bounds can be determined from the region that the data was observed from. If this region is a 50 km × 50 km square, a plausible interval of values for ρ could be $[1\text{ km}, 30\text{ km}]$. Similarly, bounds for the marginal variance can be determined by considering the scale and the amount of variation in the observed data.

In S-ANISO and K-ANISO, the priors for w_x and w_y can be controlled in a similar manner. The quantity $\sqrt{1 + w_x^2 + w_y^2}$ is the strength of the anisotropy, i.e., the ratio between the range in the longest and shortest direction. Note that $\sqrt{1 + w_x^2 + w_y^2} \geq 1$, with equality only for the baseline case $w_x = w_y = 0$. Since the priors for w_x and w_y are both Gaussian with mean 0 and the same standard deviation, $\sqrt{w_x^2 + w_y^2}$ follows a Rayleigh distribution. We can then ensure that

$$\mathrm{P}\left(\sqrt{1 + w_x^2 + w_y^2} < w_{\mathrm{strength}}\right) = 1 - \alpha$$

[3] We use the convention $\mathrm{P}(|Z| < z_{\alpha/2}) = 1 - \alpha$, when $Z \sim \mathcal{N}(0, 1)$.

by letting $v_{\boldsymbol{w}} = \sqrt{w_{\text{strength}}^2 - 1}/r_{1-\alpha} = \sqrt{-(w_{\text{strength}}^2 - 1)/2\ln(\alpha)}$. Here r_α is the quantile function of the standard Rayleigh distribution, so that $\mathrm{P}(R < r_\alpha) = \alpha$ for standard Rayleigh distributed R. We consistently use $\alpha = 0.05$ for the above priors, unless explicitly stated otherwise.

The approaches described above can also be applied to the corresponding parameters $\rho_0, w_{x,0}, w_{y,0}$ and σ_0 in the non-stationary parametrizations. Choosing the coefficient standard deviation v_{NS} is more difficult, and cannot be done in the systematic way described for the intercepts. We instead demonstrate the effects of prior width on both estimated parameters and prediction performance in Section 5.5, and use this as a guideline when choosing the standard deviation.

3.4 Discussion

Two approaches for specifying non-stationary covariance structures have been described. In the kernel convolution-based approach, the covariance structure is controlled indirectly by spatially varying kernels. The covariance between any two locations depends only on their positions and the value of the functions $\boldsymbol{\Sigma}(\cdot)$ and $\sigma(\cdot)$ in the locations, leading to a global specification of anisotropy. In addition, a closed-form expression for the covariance function is known. The SPDE-based approach gives a non-stationary GRF where the spatially varying coefficients affect the correlation structure in a local manner, but without a known closed form for the correlation function. Additionally, the correlation between two locations depends on the behavior of $\kappa(\cdot)$ and $\mathbf{H}(\cdot)$ in the region between the locations, meaning that the anisotropy is determined by local properties. As a result, the correlation structures attained from each approach are qualitatively different. This is best demonstrated with an example.

Example 3.1 (Comparison between SPDE- and kernel-based model for barrier of short range)**.** We consider the region $\mathcal{D} = [-10, 10]^2$ with a 300×300 grid. Using the SPDE-based model S-NS1 and kernel-based model K-NS, we let their matrix functions be on the form

$$\mathbf{H}(\boldsymbol{s}) = \boldsymbol{\Sigma}(\boldsymbol{s}) = \frac{1}{8}\rho^2(\boldsymbol{s})\mathbf{I}_2, \quad \boldsymbol{s} \in \mathcal{D},$$

where the approximate effective range function $\rho(\cdot)$ is given by

$$\log\left(\rho(x,y)\right) = \log(5) - 6\exp\left(-20x^2\right).$$

A plot of $\rho(\cdot)$ is shown in Figure 3.2a. The function is constant and equal to 5 in most of $\mathcal{D}$, except for a thin barrier around $x = 0$, where it rapidly decreases to a minimum value of 0.012. As a result, the GRFs obtained from both models are isotropic at a certain distance from the barrier, but become increasingly nonstationary near the barrier. Figure 3.3 shows the correlation between the location $(-1, 0)$ and the rest of $\mathcal{D}$ for both models. In the kernel-based approach, the correlation structure is indistinguishable from an isotropic correlation structure, except for a strip along the barrier where it vanishes to 0. The correlations on the right side of the barrier are unaffected by the region with short range. The correlation structure obtained with the SPDE approach is very close to isotropic on the left-hand side of the barrier, with some deformation apparent. Along the barrier and on the right-hand side, however, the correlation vanishes completely. In this case, the barrier effectively divides the region into two independent subregions.

The correlation structure obtained with the SPDE approach can be considered more natural, as the correlation decreases monotonically with distance along any direction from $(-1, 0)$. This is consistent with the intuition that the degree of dependence between locations should decrease with distance. The effect of having regions with very short range is investigated in Bakka et al. (2019), where "barriers" of short range are exploited to get complex correlation structures.

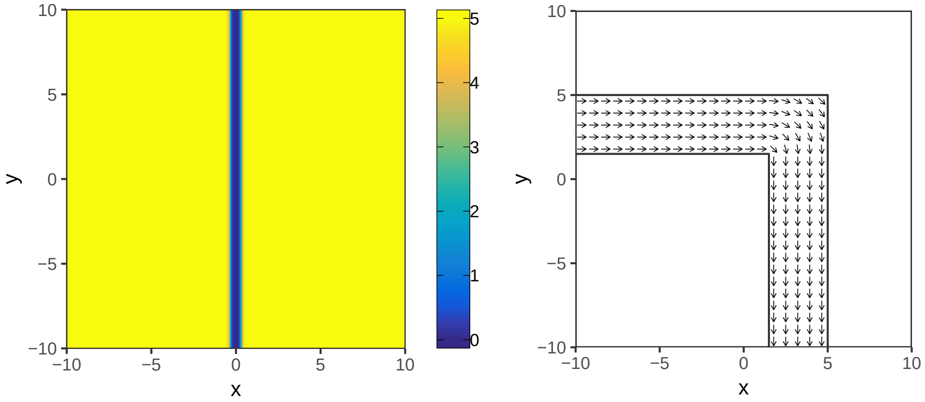
There are examples of situations where the kernel-based model can be more appropriate. Teleconnections are a phenomenon where climate and weather anomalies are connected over large distances across the globe. The El Niño-Southern Oscillation, which is investigated in Diaz et al. (2001), is an example of this. When modeling such processes, the ability to have non-monotonic correlation structures might be advantageous. △

We see that having the same effective range function $\rho(\cdot)$ in the SPDE- and kernel-based models does not lead to the same GRF. The same holds for the vector field $\boldsymbol{w}(\cdot)$. To demonstrate, we consider another example.

Example 3.2 (Comparison between SPDE- and kernel-based model with varying direction of longest range)**.** We consider the models S-NS1 and K-NS on the region $\mathcal{D} = [-10, 10]^2$ with a 300×300 grid, and let

$$\mathbf{H}(\boldsymbol{s}) = \boldsymbol{\Sigma}(\boldsymbol{s}) = \frac{1}{8}\left(\mathbf{I}_2 + \boldsymbol{w}(\boldsymbol{s})\boldsymbol{w}(\boldsymbol{s})^{\mathsf{T}}\right), \quad \boldsymbol{s} \in \mathcal{D},$$

where $\boldsymbol{w}(\cdot)$ is shown in Figure 3.2b. In Figure 3.4, we focus on the region $[-5, 5]$ and compare the correlation structure obtained from both approaches, by showing level curves around three locations. The level curves centered in $(-3.25, 3.25)$ and $(3.25, -3.25)$ are practically indistinguishable for the two approaches, as $\boldsymbol{w}(\cdot)$ is constant in the vicinity around both locations. Around $(3.25, 3.25)$, however, the correlation structures are clearly different. While S-NS1 leads to a smooth correlation structure that follows along the changing direction of $\boldsymbol{w}(\cdot)$, K-NS has more irregular level curves. △



(a) $\rho(\cdot)$

(b) $\boldsymbol{w}(\cdot)$

Figure 3.2: (a) Function $\rho(\cdot)$ used in Example 3.1. The minimum value is 0.012. (b) The vector field $\boldsymbol{w}(\cdot)$ from Example 3.2. The length of the arrows have been scaled by a factor of 0.12, and $\boldsymbol{w}(\cdot)$ is zero outside the passage.

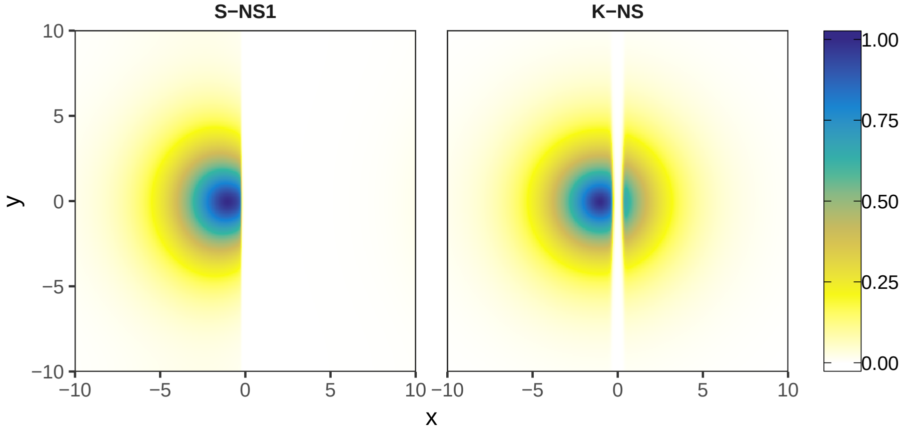
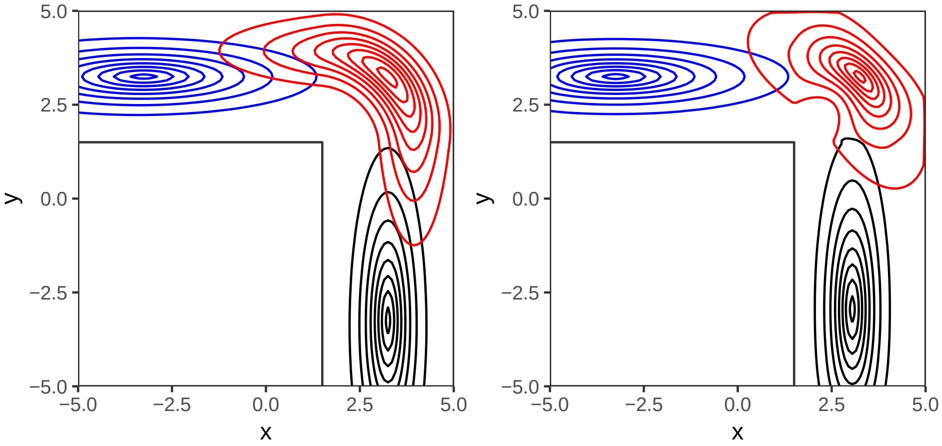


Figure 3.3: Resulting correlation structure for models S-NS1 and K-NS in Example 3.1.



(a) SPDE model S-NS1

(b) Kernel model K-NS

Figure 3.4: Comparison of correlation structures from Example 3.2. The correlation structures centered in locations $(-3.25, 3.25)$ (——), $(3.25, 3.25)$ (——) and $(3.25, -3.25)$ (——) are shown, represented by level curves on 8 levels evenly spaced between 0.14 and 0.95.

Based on the above examples, it is important to note that, while we use the same notation to describe the decompositions of $\mathbf{H}(\cdot)$ and $\mathbf{\Sigma}(\cdot)$, they do not affect the accompanying models in the same way.

For the SPDE-based approach, the marginal variance and correlation structure are both coupled to the coefficients $\kappa^2(\cdot)$ and $\mathbf{H}(\cdot)$. Three non-stationary parametrizations are proposed, each attempting to control the marginal variance by using the proposed approximations. The model S-NS1E controls the marginal variance exactly, but relies on the computation of the partial inverse of $\mathbf{Q_H}$. This is done by the function `inla.qinv` in the R-INLA package. In Figure 3.5, the time needed to compute the partial inverse is shown for different grid sizes $n = n_x n_y$. As n increases, we see that the run-time of the function grows as $\mathcal{O}(n^{1.5})$. The non-linear time complexity makes the execution of `inla.qinv` a computational bottleneck when S-NS1E is used with a large grid. In Fuglstad and Castruccio (2020), `inla.qinv` is utilized successfully to control the marginal variance of an SPDE model with a grid size of $n = 15392$. For situations where n is larger than this, and the model must be fitted repeatedly (e.g. cross-validation), S-NS1E becomes very time-consuming. In the kernel-based approach, the marginal variance is separated from the correlation structure, and the non-stationary parametrization K-NS leads to exact control of the marginal variance.

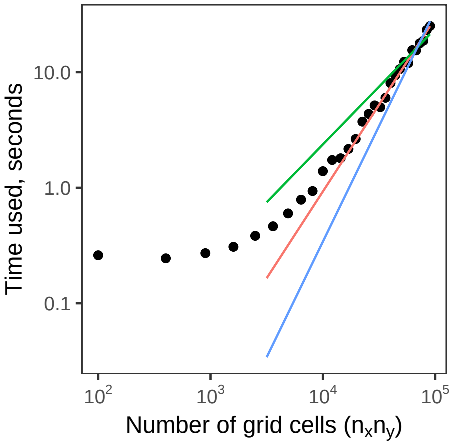


Figure 3.5: Average runtime of `inla.qinv` based on 10 runs. In addition to the data, we show lines satisfying $y \propto x$ (—), $y \propto x^{1.5}$ (—), and $y \propto x^2$ (—).

When a covariate-based parametrization such as S-NS1 or K-NS is used, there is a practical difference between the two approaches. For K-NS, we only need the value of the covariate in the observation and prediction locations. In the SPDE-based approach, setting up the model requires the value of the covariate over the entire grid. If the covariate is some physical property like elevation or land type, and not simply a known function like $\boldsymbol{x}(\boldsymbol{s}) = \|\boldsymbol{s}\|$, this could be problematic. For many situations, obtaining the value of such a covariate over a large grid is infeasible. A possible solution is to interpolate/extrapolate the value of the covariate based on its known values.

Chapter 4

Models and inference

In this chapter we describe the model that is used for inference on spatial processes in Chapters 5 and 6. We take the Bayesian approach to modeling, and describe two tools for performing both inference and prediction. We also discuss a known issue with one of the tools, which is demonstrated by performing inference on simulated data.

4.1 Model and priors

Our focus is on modeling processes that occur in the plane. Let $\eta(\cdot)$ be such a process, defined on some region $\mathcal{D} \subset \mathbb{R}^2$. We model $\eta(\cdot)$ by decomposing it into multiple terms, leading to

$$\eta(\boldsymbol{s}) = \mu + \boldsymbol{x}(\boldsymbol{s})^\mathsf{T}\boldsymbol{\beta} + u(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D}.$$

The components are the intercept $\mu \in \mathbb{R}$, the linear effect $\boldsymbol{x}(\boldsymbol{s})^\mathsf{T}\boldsymbol{\beta}$ consisting of the spatially varying covariates $\boldsymbol{x}(\boldsymbol{s}) \in \mathbb{R}^p$ and the coefficient of the effect $\boldsymbol{\beta} \in \mathbb{R}^p$, and a GRF $u(\cdot)$ intended to capture the residual dependencies between nearby locations. We model $u(\cdot)$ using one of the parametrizations from Chapter 3. When observing the process in a location $\boldsymbol{s} \in \mathcal{D}$, we obtain the noisy measurement y satisfying

$$y \mid \eta(\boldsymbol{s}), \sigma_\varepsilon \sim \mathcal{N}(\eta(\boldsymbol{s}), \sigma_\varepsilon^2).$$

Here, σ_ε is the standard deviation of the measurement error. If we observe $\eta(\cdot)$ in multiple locations $\boldsymbol{s}_1, \dots, \boldsymbol{s}_n \in \mathcal{D}$, we then obtain the observed data $\boldsymbol{y} = (y_1, \dots, y_n)$ with y_i observed at $\boldsymbol{s}_i$ for $i = 1, \dots, n$. The vector $\boldsymbol{y}$ satisfies

$$\boldsymbol{y} \mid \boldsymbol{\eta}, \sigma_\varepsilon \sim \mathcal{N}_n(\boldsymbol{\eta}, \sigma_\varepsilon^2 \mathbf{I}_n),$$

where $\boldsymbol{\eta} = \mu \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{u}$. Here, $\mathbf{1}_n$ is an n-dimensional vector of ones, $\mathbf{X}$ is a $n \times p$ matrix with i-th row $\boldsymbol{x}(\boldsymbol{s}_i)^\mathsf{T}$, and $\boldsymbol{u} = (u_1, \ldots, u_n)$ with $u_i = u(\boldsymbol{s}_i)$ for $i = 1, \ldots, n$.

We use a centered Gaussian prior for the intercept and the linear effect coefficients:

$$\mu \sim \mathcal{N}(0, v_\mu^2), \quad \boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{0}, v_{\boldsymbol{\beta}}^2 \mathbf{I}_p).$$

The prior setup for the spatial effect $u(\cdot)$ is described in Section 3.3, and depends on which parametrization is used. The measurement error standard deviation σ_ε is equipped with the penalized complexity (PC) prior, which is introduced in Simpson et al. (2017). This leads to an exponential prior for the standard deviation, which is specified by choosing a bound U and weight α so that $\mathrm{P}(\sigma_\varepsilon > U) = \alpha$.

Since the priors for μ and $\boldsymbol{\beta}$ are Gaussian, this is a *latent Gaussian model.* We refer to $\mu, \boldsymbol{\beta}$, and $\boldsymbol{u}$ as the latent field of the model, and collect these in the vector $\boldsymbol{\xi}$. The parameters of the model are stored in the vector $\boldsymbol{\theta}$, which contains σ_ε and the parameters controlling the spatial effect $u(\cdot)$.

4.2 Bayesian inference

In this thesis, we work within the Bayesian framework for inference. Given the observed data $\boldsymbol{y}$, we are mainly interested in doing two things:

- Obtaining posterior distributions of the latent field $\boldsymbol{\xi}$ and the hyperparameters $\boldsymbol{\theta}$, i.e., estimating the posterior marginals $p(\xi_i|\boldsymbol{y})$ and $p(\theta_i|\boldsymbol{y})$.
- Predicting the process $\eta(\cdot)$ in unobserved locations, based on the observed data $\boldsymbol{y}$, i.e., estimating the posterior marginals $p(\eta(\boldsymbol{s}^*)|\boldsymbol{y})$ and $p(y^*|\boldsymbol{y})$, where y^* is a new observation at $\boldsymbol{s}^*$.

Two different computational approaches are utilized for performing inference, depending on whether the model for $u(\cdot)$ is SPDE- or kernel-based. For the SPDE-based models, inference is done using `R-INLA`, while `BayesNSGP` is used for the kernel-based models.

R-INLA

`R-INLA` is an R package for performing Bayesian inference with latent Gaussian models. While MCMC estimates posterior distributions by generating samples, `R-INLA` makes use of the Integrated Nested Laplace Approximations (INLA) methodology (Rue et al., 2009), which uses numerical integration to directly approximate the posterior distributions. `R-INLA` offers a plethora of alternatives to choose from when it comes to likelihoods, priors and latent effects. In the context of our model, INLA requires that $\boldsymbol{u}$ is a GMRF. Since this is ensured with SPDE-based spatial effects, R-INLA is an appropriate tool for inference.

While the covariate-based SPDE models described in Section 3.1 have not been implemented in R-INLA, custom models can be specified through the function inla.rgeneric.define. Using the recipe outlined in the rgeneric-vignette, we implement model definitions for S-ISO, S-ANISO, S-NS1, S-NS1E, and S-NS2. These definitions contain all necessary components for specifying a GMRF in R-INLA, such as the construction of the precision matrix $\mathbf{Q}$, the log-density function of the priors and the logarithm of the normalizing constant.

The construction of $\mathbf{Q}$, which is outlined in Fuglstad et al. (2015a), is particularly technical. Since $\mathbf{Q}$ is a sparse matrix, dedicated data structures must be used for storing it. For this purpose, we use the function sparseMatrix from the library Matrix. This function allows us to construct $\mathbf{Q}$ by specifying the value and position of each non-zero element in the matrix. Constructing $\mathbf{Q}$ involves performing calculations in each cell of the grid. The most straightforward way to do this is to iterate through the grid using a nested for-loop. However, for-loops are slow in R, especially when compared to compiled languages like C++. For big grid sizes, this slows things down considerably, as the precision matrix must be constructed many times during the inference. The construction has therefore been implemented using only vectorized R operations.

In the non-stationary parametrizations, the construction of $\mathbf{Q}$ depends on spatial covariates. This is done by specifying design matrices for the functions that are modeled through linear regression, i.e., $\mathbf{X}_\rho$, $\mathbf{X}_{w_x}$, $\mathbf{X}_{w_y}$, and $\mathbf{X}_\sigma$. This approach allows for quick computation of the corresponding functions for different regression coefficients. For each parametrization, the prior setup had to be implemented manually. In R-INLA, priors are specified through their density function. Since all of the SPDE model parameters have Gaussian priors, the priors can be implemented using only the dnorm function in R.

After creating the model definitions, we can use inla.rgeneric.define to initiate a model:

```
custom.model = inla.rgeneric.define(model = model.definition,...).
```

Parameters necessary for setting up a model are specified in (...). Such parameters can be prior.HP, which is a list specifying the prior hyperparameters, or X.rho, X.wx, X.wy, and X.sigma, which are covariate design matrices in the non-stationary parametrizations.

BayesNSGP

The BayesNSGP package, which is outlined in Risser and Turek (2019), is dedicated specifically to performing non-stationary spatial modeling with the kernel-based approach discussed in Section 3.2. The package offers off-the-shelf functionality for performing Bayesian inference using the MCMC methodology, and includes the approximate Gaussian likelihood described in Section 2.3 for dealing with large spatial datasets.

While the package offers plenty of choices for controlling both $\sigma(\cdot)$ and $\mathbf{\Sigma}(\cdot)$, none of the parametrizations K-ISO, K-ANISO, and K-NS are available. In addition, there is no functionality that allows the user to directly specify custom models and parametrizations. As a result, performing inference with the desired parametrizations proved difficult. After familiarizing ourselves with source code of the package and understanding how the already existing parametrizations had been implemented, we were able to add new parametrizations by directly modifying the source code and rebuilding the package.

A number of the hardcoded prior choices in `BayesNSGP` are questionable. Most notably, both the marginal variance of the GRF and the variance of the observational error are given uniform priors on the interval $[0, a]$ for some specified a. If we consider the marginal variance, the prior density function satisfies $\pi(\sigma^2) \propto \mathbb{1}(0 \leq \sigma^2 \leq a)$. It can then be shown that the prior density for the marginal standard deviation σ satisfies $\pi(\sigma) \propto \mathbb{1}(0 \leq \sigma \leq \sqrt{a})\sigma$. This is unfortunate, as the density increases with the value of σ, and leads to a bias for higher values of σ. As a result, the value of σ is consistently overestimated. The exact same argument applies for the variance and standard deviation of the observational error. We therefore add the ability to have a log-normal prior for the marginal standard deviation and a PC prior for the precision of the measurement error, as desired.

4.3 Posterior multimodality

As discussed in Rue et al. (2009), one of the main issues with INLA is that only unimodal posterior distributions can be represented accurately. When the true posterior distribution of the parameters $\boldsymbol{\theta}$ or the latent field $\boldsymbol{\xi}$ is multimodal, a unimodal approximation is obtained. In essence, INLA relies on an approximation obtained by centering a Gaussian distribution at the mode of the posterior distribution and matching the curvature at the mode. The mode of the posterior distribution is determined by a Newton-Raphson iteration. For densities on the form

$$p(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^{\mathsf{T}}\mathbf{Q}\boldsymbol{x} + \sum_{i=1}^{n} g_i(x_i)\right), \tag{4.1}$$

this entails choosing some initial guess $\boldsymbol{\mu}^{(0)}$ of the mode, and performing a second-order expansion of each $g_i(\cdot)$ around $\mu_i^{(0)}$,

$$g_i(x_i) \approx g_i(\mu_i^{(0)}) + b_i x_i - \frac{1}{2} c_i x_i^2.$$

By replacing each $g_i(\cdot)$ in Equation (4.1) with this expansion, we obtain a Gaussian approximation with precision matrix $\mathbf{Q} + \operatorname{diag}(\boldsymbol{c})$ and mode $\boldsymbol{\mu}^{(1)}$ satisfying the equation $(\mathbf{Q} + \operatorname{diag}(\boldsymbol{c}))\boldsymbol{\mu}^{(1)} = \boldsymbol{b}$. Here, $\boldsymbol{b} = (b_1, \ldots, b_n)$ and $\boldsymbol{c} = (c_1, \ldots, c_n)$. Using $\boldsymbol{\mu}^{(1)}$ as our new guess we can repeat the procedure, and this is done until convergence.

When there are multiple modes, only one is identified by the above iteration, and the approximation to the posterior becomes unimodal. This issue does not occur with MCMC, where samples are generated from the true posterior distribution. However, it can still be challenging to sample from a multimodal posterior with MCMC. Slow mixing is the most common issue, where the chain struggles to move between modes that are separated by a low probability region.

With this in mind, the decompositions for $\mathbf{H}(\cdot)$ in Equations (3.9), and $\boldsymbol{\Sigma}(\cdot)$ in Equation (3.13), are problematic, as the vector components are non-identifiable. For example, exchanging $\boldsymbol{w}(\cdot)$ with $-\boldsymbol{w}(\cdot)$ leads to an identical model. Let the components $w_x(\cdot)$ and $w_y(\cdot)$ be modeled by the regressions

$$\begin{aligned} w_x(\boldsymbol{s}) &= w_{x,0} + \boldsymbol{x}_{w_x}(\boldsymbol{s})^\mathsf{T} \boldsymbol{\beta}_{w_x}, \\ w_y(\boldsymbol{s}) &= w_{y,0} + \boldsymbol{x}_{w_y}(\boldsymbol{s})^\mathsf{T} \boldsymbol{\beta}_{w_y}, \end{aligned}$$

with a zero-mean Gaussian prior for $\boldsymbol{\theta_w} = (w_{x,0}, w_{y,0}, \boldsymbol{\beta}_{w_x}, \boldsymbol{\beta}_{w_y})$. Then, the posterior distribution for the parameters is bimodal or worse: for any value of $\boldsymbol{\theta_w}$ and observed data $\boldsymbol{y}$, the posterior satisfies $p(\boldsymbol{\theta_w}|\boldsymbol{y}) = p(-\boldsymbol{\theta_w}|\boldsymbol{y})$. Since we perform inference with INLA for the SPDE-based models and MCMC for the kernel-based models, this could potentially introduce unforeseen differences when comparing the two approaches. To get an idea of the difference between the true and approximate posterior density, we explore the issue in an example. In the example, we generate data from both an isotropic and a geometrically anisotropic process. Based on the observed data, we perform inference using the SPDE-based S-ANISO with `R-INLA`, and the kernel-based K-ANISO with `BayesNSGP`. We then compare obtained posterior distributions from each approach.

Example 4.1 (Comparison of estimated anisotropy from INLA and MCMC)**.** Let the region of interest be $\mathcal{D} = [-5, 5]^2$, and consider the processes

$$\psi_j(\boldsymbol{s}) = 1 + \omega_j(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D},$$

for $j = 1, 2$. Here, $\omega_1(\cdot)$ is a centered, isotropic GRF with effective range 1.5 and marginal variance 1. $\omega_2(\cdot)$ is a centered, geometrically anisotropic GRF with marginal variance 1, shortest and longest effective ranges 1.5 and 6, and longest range in the direction of $(1,1)$.

For $j = 1, 2$, we observe the value of $\psi_j(\cdot)$ with measurement error in 50 uniformly sampled locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{50} \in \mathcal{D}$, leading to observations $y_i = \psi_j(\boldsymbol{s}_i) + \varepsilon_i$ for $i = 1, \ldots, 50$, where $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.1^2)$. Based on the observed data, we perform inference with the model described in Section 4.1. We let $\eta(\cdot)$ consist of an intercept and a spatial effect, i.e., $\eta(\boldsymbol{s}) = \mu + u(\boldsymbol{s})$ for $\boldsymbol{s} \in \mathcal{D}$. Inference is performed with $u(\cdot)$ specified by both the SPDE-based S-ANISO and the kernel-based K-ANISO.

We are interested in the joint posterior distribution of the parameters (w_x, w_y). While INLA does not return an estimate of the joint posterior density of the model parameters, the function `inla.hyperpar.sample` allows us to generate samples from the approximate joint distribution of w_x and w_y.

In Figure 4.1 we show the estimates of the joint posterior densities of w_x and w_y from both INLA and BayesNSGP. The samples from BayesNSGP are represented as a hexagonal bin plot, where the fill color of each tile indicates the proportion of samples contained inside the tile. We also show estimated density level curves based on both the samples from BayesNSGP and INLA. In addition, the marginal densities of w_x and w_y are added opposite to the corresponding axis, with color indicating method.

When the data comes from the isotropic process $\psi_1(\cdot)$, the true joint posterior distribution is bimodal and clearly non-Gaussian, as the level curves are distinctly non-elliptic. The level curves from INLA, however, are close to elliptic and indicate that the approximated joint posterior is unimodal and more regular in shape. The mode identified by INLA misses somewhat along the x-direction, and the distribution as a whole underrepresents the uncertainty of the true joint posterior.

With data from the geometrically anisotropic process $\psi_2(\cdot)$, the true joint distribution of w_x and w_y is smoother, and has two modes that are clearly separated. In this case, INLA does a better job at approximating the shape of the distribution. It correctly identifies the mode in the third quadrant, but misses its location somewhat. The approximate distribution is also more concentrated around the mode than the true distribution, which leads to a lower uncertainty in the estimates of both parameters.

Its hard to anticipate how these differences affect the predictions attained from each approach, and whether they have an impact on the predictive performance. In this particular example we have also predicted the value of the processes in 950 uniformly sampled locations. When predicting the anisotropic process, S-ANISO leads to a mean CRPS of 0.315 and an RMSE of 0.579, while K-ANISO results in 0.318 and 0.586. Corresponding values for the isotropic process are 0.508 and 0.895 for S-ANISO, and 0.516 and 0.912 for K-ANISO.

In both cases INLA does a marginally better job despite the issues described. While a single example is insufficient for drawing any reliable conclusions, this indicates that, for the stationary models, the differences between the SPDE-based approach in INLA and kernel-based approach in BayesNSGP are negligible. △

For practical reasons, we are interested in extracting quantities such as the posterior mean and median of w_x and w_y from the joint posterior distributions. For the results obtained from INLA, both of these are computed internally and are easily available. The MCMC samples of w_x and w_y from BayesNSGP will often have bimodality present. This bimodality combined with the symmetry around the origin leads to estimated posterior means and medians close to 0. We therefore use the location of the modes as our posterior estimates of w_x and w_y. In order to extract the location of these, we use the package `ContaminatedMixt` to fit bivariate Gaussian mixture models to the samples. When the samples are unimodal, we fit a mixture consisting of a single group. For samples with bimodality present, we fit a mixture with two groups and use one of the estimated modes as our estimate.

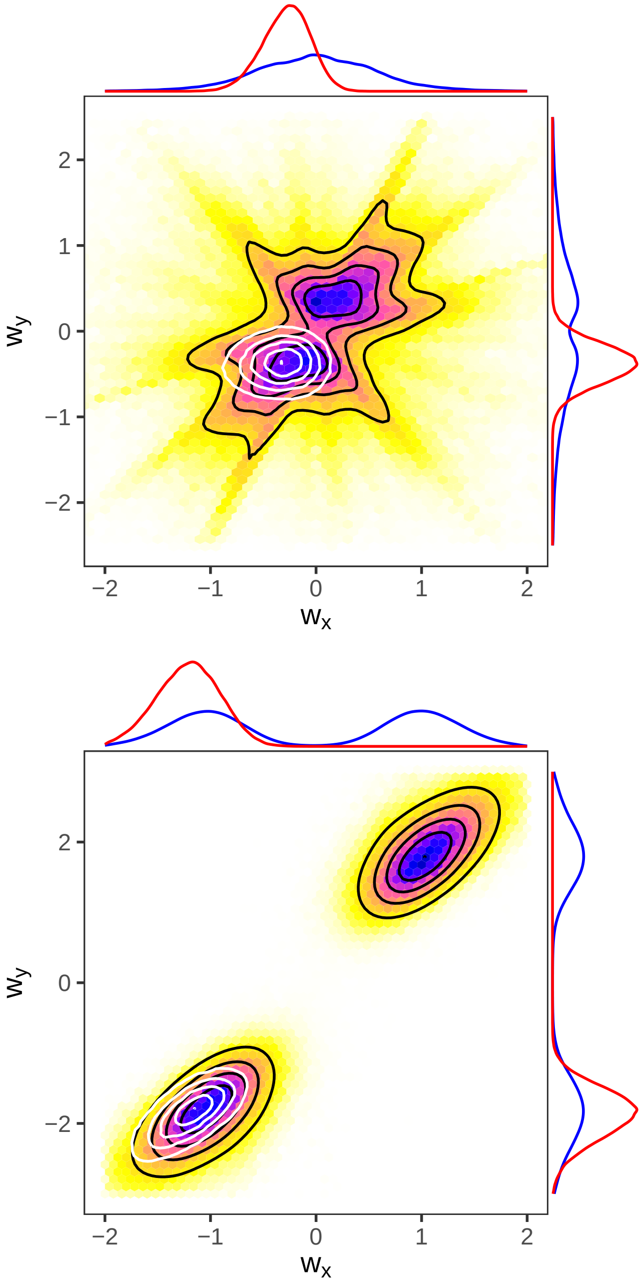


Figure 4.1: Comparison of the joint posterior distribution of w_x and w_y from Example 4.1, for both isotropic (top) and anisotropic (bottom) data. The MCMC samples from BayesNSGP are shown as a hexagonal bin plot, where a darker color indicates higher density. The black and white lines are estimated density level curves based on the samples from BayesNSGP and INLA, respectively. In addition, we show the marginal densities of w_x and w_y opposite the x- and y-axis for BayesNSGP (—) and INLA (—).

Chapter 5

Simulation study

In this simulation study we generate data from known spatial processes and perform inference on the data with several candidate models from Chapter 3. The goal is to get an understanding of under which circumstances a model performs well, with respect to both prediction scores, and in the estimation of central features, such as the correlation structure and marginal variance.

We generate the data from four GRF models, in which two are stationary and two are non-stationary. We refer to the resulting studies as Studies 1, 2, 3, and 4. Both the GRFs and the results obtained with the candidate models are presented, and the chapter ends with a discussion of the results. We start by describing the general setup of the study.

5.1 Study setup

Our aim is to construct a set of simulation studies that consists of realistic scenarios. Given a spatial process over a region, this typically entails observing its value only in certain irregular locations, usually corresponding to measurement stations. Therefore, we generate realizations from a spatial process and observe its value only in a fixed number of uniformly sampled locations. For each study, the observed process is on the form

$$\psi(\boldsymbol{s}) = 1 + \omega(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D}, \tag{5.1}$$

where $\omega(\cdot)$ is a centered GRF representing a spatial effect and $\mathcal{D}$ is the rectangle $[-5, 5]^2$. The covariance structure of $\omega(\cdot)$ is specified in each study. While it is common to let the mean depend on covariates, this leads to a non-stationary mean structure. Furthermore, it is impossible to separate a non-stationary covariance structure from a non-stationary mean structure given only a single realization (Gelfand et al., 2010, page 30). As our focus is on non-stationarity in the covariance structure, we therefore let the mean be stationary throughout this simulation study. For all studies, we use the GMRF approximation from Section 3.1 to generate realizations from $\omega(\cdot)$. This allows us to quickly obtain entire realizations of $\omega(\cdot)$ even for large grid sizes. We avoid boundary issues by generating the realizations on a rectangle larger than $\mathcal{D}$.

We observe the value of $\psi(\cdot)$ with measurement error in 1000 uniformly sampled locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{1000}$, leading to the pairs $(y_i, \boldsymbol{s}_i)_{i=1}^{1000}$ where $y_i = \psi_i + \varepsilon_i$. Here $\psi_i = \psi(\boldsymbol{s}_i)$ and $\varepsilon_1, \ldots, \varepsilon_{1000} \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.1^2)$ is unstructured noise representing measurement error. The first m_{inf} pairs are used for the model inference. The values of the process in the remaining $m_{\text{pred}} = m - m_{\text{inf}}$ locations are predicted, and compared with the true, observed values using the prediction scoring rules described in Section 2.4. The values of m_{inf} and m_{pred} are specified in each study, and we increase the value of m_{inf} as the process $\omega(\cdot)$ becomes more complex. Note that the value of the structured component ψ_i is predicted, not y_i. When evaluating the predictions, we therefore use ψ_i as the true values.

This procedure of simulating observations, estimating the models and predicting unobserved locations is replicated 20 times in each study, using different data and observation locations each time. While 20 replications are insufficient for making any definite conclusions, it is sufficient for getting an understanding of how the different models perform.

The model used for the inference is the LGM described in Section 4.1. We let $\eta(\cdot)$ consist of only an intercept μ and a spatial effect $u(\cdot)$, so that

$$\eta(\boldsymbol{s}) = \mu + u(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D},$$

and the response at location $\boldsymbol{s}_i$ satisfies $y_i \mid \eta(\boldsymbol{s}_i), \sigma_\varepsilon \sim \mathcal{N}(\eta(\boldsymbol{s}_i), \sigma_\varepsilon)$ for $i = 1, \ldots, 1000$, where σ_ε is the standard deviation of the measurement error. For each study, we perform inference using five candidate models for $u(\cdot)$: The SPDE-based S-ISO, S-ANISO, S-NS1, and S-NS2, and the kernel-based K-NS. These models are presented in Chapter 3. There is an imbalance, as we use four SPDE-based models and only one from the kernel-based approach. While we could also include the stationary K-ISO and K-ANISO, these lead to the same covariance structures as S-ISO and S-ANISO, respectively. For this simulation study, we therefore only include the stationary models from the SPDE-based approach.

In order to minimize the effects from the periodic boundary conditions, the inference with the SPDE-based models is done on $[-10, 10]^2$ with a 200×200 grid. This leads to a 100×100 grid on the region of interest, $\mathcal{D}$. For K-NS we use an exact Gaussian likelihood, and generate 50000 posterior samples of the model parameters. While the MCMC chains usually converge well within the first 500 samples, we dismiss the first 5000 as burn-in to be certain. Out of the remaining 45000, we use 5000 thinned samples for generating the posterior predictions.

Ideally, more samples should be generated for both the model estimation and the prediction, but we limit ourselves due to time constraints.

In K-NS, we let the marginal standard deviation $\sigma(\cdot)$ be spatially constant. The same holds for the approximation $\sigma(\cdot)$ in S-NS1 and S-NS2. In general, the latter two models lead to a marginal standard deviation that is not constant. When a single value of the marginal standard deviation is mentioned in the context of these two models, we have computed the partial inverse of the precision matrix $\mathbf{Q}$ and used the diagonal elements to compute a spatial average standard deviation over $\mathcal{D}$.

In S-NS1, S-NS2, and K-NS, the functions $\log(\rho(\cdot)), w_x(\cdot)$, and $w_y(\cdot)$ are modeled using regression on spatial covariates. For $\phi(\cdot) \in \{\log(\rho(\cdot)), w_x(\cdot), w_y(\cdot)\}$, we let $\phi(\cdot)$ consist of an intercept and a single covariate effect:

$$\phi(\boldsymbol{s}) = \phi_0 + \beta_\phi z_\phi(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D}.$$

In the cases where a "true" covariate exists, i.e., a covariate is used when generating the observed data, this is used for $z_\phi(\cdot)$. Otherwise, we let $z_\phi(\cdot)$ be the standardized x-coordinate of the location, where the standardization is done so that $z_\phi(\cdot)$ has mean 0 and standard deviation 1 over $\mathcal{D}$. The covariates used in each study is specified closer in Section 5.2.

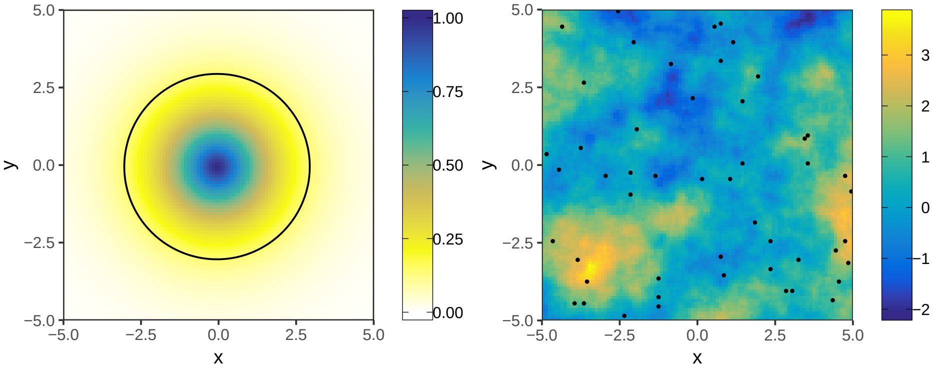
In addition to μ and σ_ε, the models will have an additional number of parameters depending on which parametrization is used for $u(\cdot)$. The covariance structures of the stationary models S-ISO and S-ANISO are specified by 2 and 4 parameters, respectively. The non-stationary models S-NS1, S-NS2, and K-NS each contain 7 parameters in total.

5.2 Study designs

Here we present the choices that are particular to each Study, such as the values of m_{inf} and m_{pred}, and the GRF $\omega(\cdot)$ from Equation (5.1). For consistency, we let every $\omega(\cdot)$ have a constant marginal variance equal to 1. While S-NS2 in Study 4 leads to a spatially varying marginal variance, we only sample data from regions where it is constant and equal to 1. The correlation structure is demonstrated by computing the correlation between the grid cell closest to $(0, 0)$ and the rest of the grid, except in Study 3, where we show level curves centered in 3 locations.

Study 1

We use $m_{\text{inf}} = 50$ data points for inference, and the remaining $m_{\text{pred}} = 950$ for evaluating predictions. The GRF $\omega(\cdot)$ is stationary and isotropic, with an effective range of 3. In Figure 5.1 we show a realization from $\omega(\cdot)$ on a 100×100 grid,



(a) Correlation

(b) Realization

Figure 5.1: GRF $\omega(\cdot)$ used in Study 1. (a) The correlation between location $(0, 0)$ and $\mathcal{D}$, with the 0.14 level curve (—–). (b) Realization of $\omega(\cdot)$ with an example of 50 uniformly sampled locations (●). Both figures use a 100×100 grid for $\mathcal{D}$.

and a demonstration of the correlation structure. As expected, the correlation structure has circular level curves, and the realization does not seem to exhibit a stronger degree of dependency along any direction.

For S-NS1, S-NS2, and K-NS, the standardized x-coordinate is used as the covariate in $\rho(\cdot)$, $w_x(\cdot)$, and $w_y(\cdot)$.

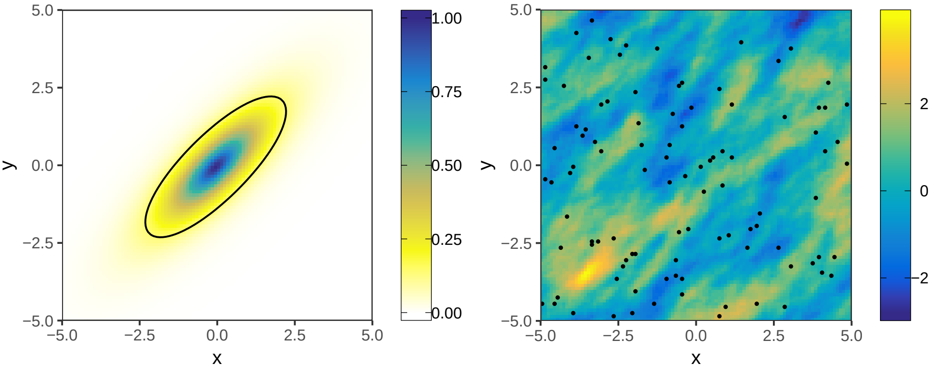
Study 2

We increase the number of observations used for inference to $m_{\text{inf}} = 100$, and use $m_{\text{pred}} = 900$ for evaluating predictions. We let $\omega(\cdot)$ be a stationary GRF with geometric anisotropy. The longest and shortest effective ranges are equal to 3 and 0.5, and the longest range is in the direction $(1, 1)$. Figure 5.2 shows a realization of $\omega(\cdot)$ and its correlation structure. The level curves of the correlation structure are elliptic, and the realization clearly has a higher degree of dependency along $(1, 1)$.

For S-NS1, S-NS2, and K-NS, the standardized x-coordinate is used as the covariate in $\rho(\cdot)$, $w_x(\cdot)$, and $w_y(\cdot)$.

Study 3

The number of locations used for inference is increased again, leading to $m_{\text{inf}} = 200$ and $m_{\text{pred}} = 800$. We generate $\omega(\cdot)$ from the non-stationary model S-NS1E from Section 3.1, which allows us to control the marginal variance exactly. For



(a) Correlation

(b) Realization

Figure 5.2: GRF $\omega(\cdot)$ used in Study 2. (a) The correlation between location $(0, 0)$ and $\mathcal{D}$, with the 0.14 level curve (——). (b) Realization of $\omega(\cdot)$ with an example of 100 uniformly sampled locations (●). Both figures use a 100 × 100 grid for $\mathcal{D}$.

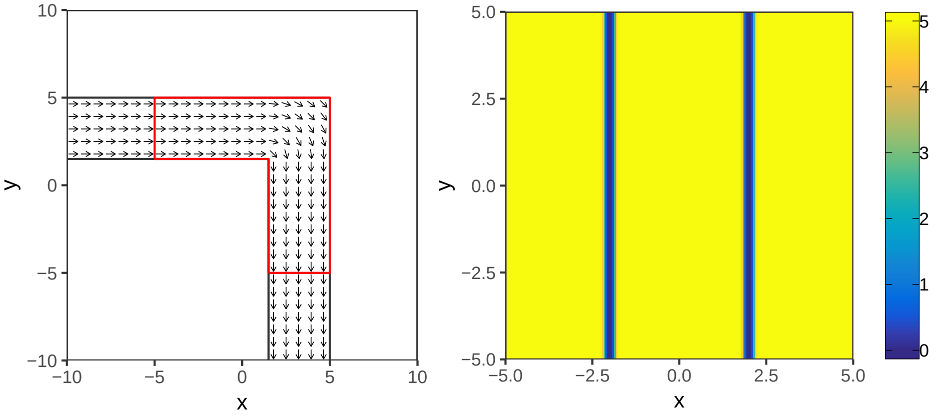
this study, we consider the same setup as in Example 3.2. We let

$$\mathbf{H}(\boldsymbol{s}) = \frac{1}{8}\rho^2(\boldsymbol{s})\left(\mathbf{I}_2 + \boldsymbol{w}(\boldsymbol{s})\boldsymbol{w}(\boldsymbol{s})^\mathsf{T}\right), \quad \boldsymbol{s} \in \mathcal{D}$$

where the baseline effective range function $\rho(\cdot)$ is spatially constant and equal to 1, while the vector function $\boldsymbol{w}(\cdot)$ is shown in Figure 5.3a. We show $\boldsymbol{w}(\cdot)$ over the entire SPDE model region $[-10, 10]^2$, and the vector field is zero outside the passage. Further, our region of interest is not $\mathcal{D}$, but rather $\mathcal{D} \setminus [-5, 1.5]^2$, which is outlined in red.

Figure 5.4a demonstrates the correlation structure around the grid cells closest to $(-3.25, 3.25), (3.25, 3.25)$ and $(-3.25, 3.25)$. Since the vector field is constant in the vicinity around $(-3.25, 3.25)$ and $(3.25, -3.25)$, the correlation structures centered in these points look nearly geometrically anisotropic. This does not hold around $(3.25, 3.25)$, where the correlation structure follows along the changing direction of the vector field. The realization shown in Figure 5.4b further underlines this, as there clearly is a stronger degree of dependency along the direction of $\boldsymbol{w}(\cdot)$.

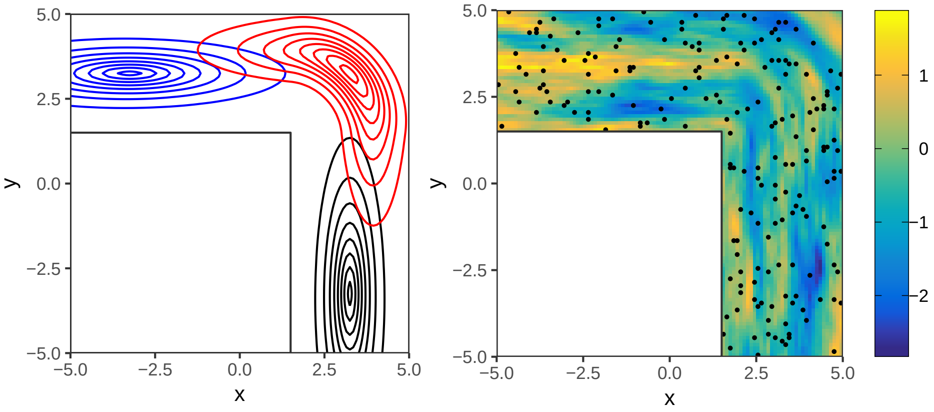
For S-NS1, S-NS2, and K-NS, the x- and the y-components of $\boldsymbol{w}(\cdot)$ are used as the covariates in $w_x(\cdot)$ and $w_y(\cdot)$, respectively, while the standardized x-coordinate is used in $\rho(\cdot)$.



(a) $\boldsymbol{w}(\cdot)$

(b) $\rho(\cdot)$

Figure 5.3: (a) The vector field $\boldsymbol{w}(\cdot)$ used in Study 3. The length of the arrows have been scaled by a factor of 0.12, and $\boldsymbol{w}(\cdot)$ is zero outside the passage. The region of interest is outlined in red. (b) The effective range process $\rho(\cdot)$ from Study 4, on a 100×100 grid. The minimum value is 0.012.



(a) Correlation

(b) Realization

Figure 5.4: GRF $\omega(\cdot)$ used in Study 3. (a) The correlation structure between the grid cells closest to $(-3.25, 3.25)$ (——), $(3.25, 3.25)$ (——) and $(3.25, -3.25)$ (——) and the rest of the passage. Level curves are shown, for eight levels evenly spaced between 0.14 and 0.95. (b) Realization of $\omega(\cdot)$ with an example of 200 uniformly sampled locations (●). Both figures use a 100×100 grid for $\mathcal{D}$.

Study 4

We use $m_{\text{inf}} = 200$ and $m_{\text{pred}} = 800$, as in Study 3. In this study we generate $\omega(\cdot)$ from S-NS2. Taking inspiration from Example 3.1, we let

$$\mathbf{H}(\boldsymbol{s}) = \frac{1}{8}\rho^2(\boldsymbol{s})\mathbf{I}_2, \quad \boldsymbol{s} \in \mathcal{D},$$

where the effective range function $\rho(\cdot)$, which is shown in Figure 5.3b, is given by $\log(\rho(\boldsymbol{s})) = \log(5) - f(\boldsymbol{s})$, where

$$f(x, y) = 6 \exp\left[-80\left(1 + \cos\left(\frac{2\pi x}{4}\right)\right)\right], \quad (x, y) \in \mathcal{D}. \tag{5.2}$$

We see that $\rho(\cdot)$ is constant and equal to 5, expect for two barriers where the value rapidly decreases to a minimum of 0.012. Based on what we observed in Example 3.1, this results in three GRFs that are almost independent. This is further demonstrated in Figure 5.4. The correlation structure is "stopped" by the barriers, and there is no dependency between locations on opposite sides. The realization looks stationary and isotropic in each of the three sub-regions.

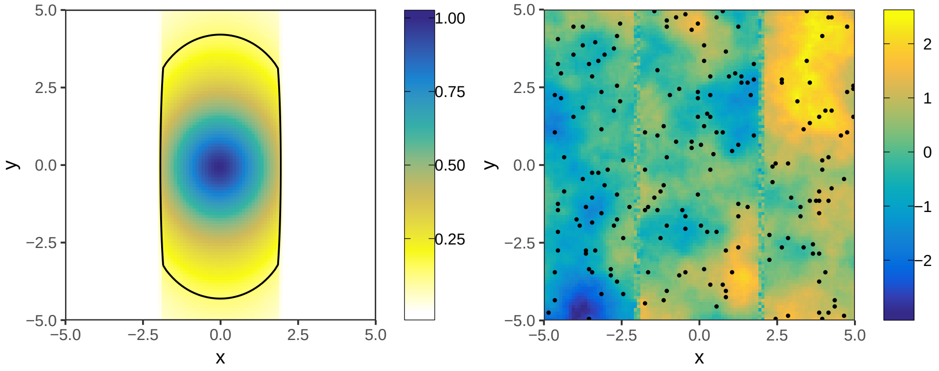
For the kernel-based K-NS, using locations too close to the barriers leads to covariance matrices that are not positive definite, during both inference and prediction. We avoid this by requiring that the observed locations have a distance of at least 0.25 from the nearest barrier. In this way, we also avoid sampling data from where the marginal variance is different from 1, which happens along the barriers.

For S-NS1, S-NS2, and K-NS, the function $f(\cdot)$ from Equation (5.2) is used as the covariate in the regression for $\rho(\cdot)$, while the standardized x-coordinate is used for $w_x(\cdot)$ and $w_y(\cdot)$.

5.3 Prior distributions

In Sections 3.3 and 4.1, the chosen families of prior distributions used for μ, σ_ε, and the parameters specifying the covariance structure of $u(\cdot)$, are described. Before inference can be done, the hyperparameters of these priors must be chosen. In a simulation study, selecting sensible values for these hyperparameters is easy, as the true values of all parameters are known. We therefore choose hyperparameters leading to wide priors that cover the true values. In this way, the true values can be recovered from the observed data, but we do not force the model to obtain the correct values.

The intercept, with true value 1, is given the prior $\mu \sim \mathcal{N}(0, 100^2)$. For the standard deviation σ_ε, with true value 0.1, we let $U = 0.5$ and $\alpha = 0.1$ in the PC prior, so that $\mathrm{P}(\sigma_\varepsilon > 0.5) = 0.1$. Next, we specify the priors for ρ, w_x, w_y, and σ



(a) Correlation

(b) Realization

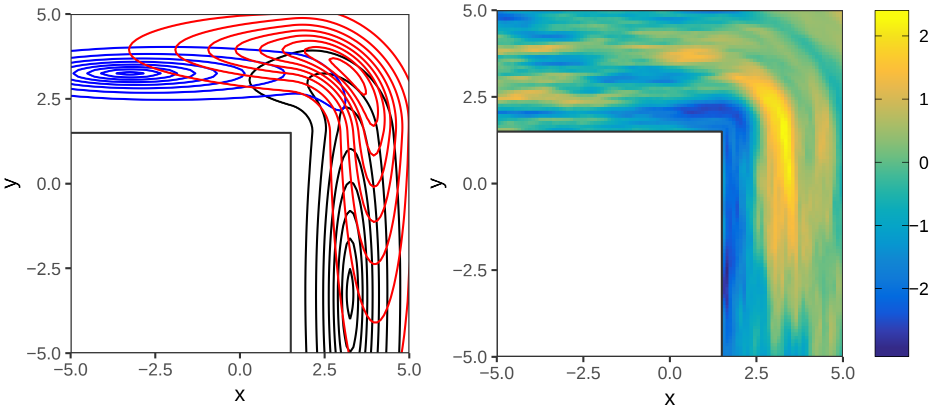
Figure 5.5: GRF $\omega(\cdot)$ used in Study 4. (a) The correlation between location $(0, 0)$ and $\mathcal{D}$, with the 0.14 level curve (——). (b) Realization of $\omega(\cdot)$ with an example of 200 uniformly sampled locations (●). Both figures use a 100×100 grid for $\mathcal{D}$.

Table 5.1: Hyperparameters for priors of covariance structure parameters. See Section 3.3 for more details.

Study	ρ, ρ_0		$w_x, w_y, w_{x,0}, w_{y,0}$	σ, σ_0		Regression coeff.
	μ_ρ	v_ρ	$v_{\boldsymbol{w}}$	μ_σ	v_σ	v_{NS}
1	0.549	1.46	1.58	-0.347	0.998	0.5
2			‖			0.5
3			‖			1.5
4			‖			1.5

in S-ISO and S-ANISO. The same priors are used for the corresponding baseline parameters in the non-stationary models, i.e., $\rho_0, w_{x,0}, w_{y,0}$, and σ_0. We follow the approach outlined in Section 3.3, with credibility level $\alpha = 0.05$.

The true values of the effective range, or baseline effective range, are 3, 1, 1, and 5 in the different studies. We let the lower and upper bound be $\rho_{\text{lower}} = 0.1$ and $\rho_{\text{upper}} = 30$. In Study 2, the anisotropy strength $\sqrt{1 + w_x^2 + w_y^2}$ is 3. Otherwise, its value is 1. The upper bound w_{strength} is therefore set to 4, so that $\mathrm{P}\left(\sqrt{1 + w_x^2 + w_y^2} > 4\right) = 0.05$. Finally, the marginal standard deviation is 1 in every study. The bounds used are $\sigma_{\text{lower}} = 0.1$ and $\sigma_{\text{upper}} = 5$. Table 5.1 shows the resulting hyperparameters.



(a) Correlation

(b) Realization

Figure 5.6: The process $\omega(\cdot)$ from Study 3 with $\beta_\rho = \beta_{w_x} = \beta_{w_y} = 1.5$. See Figure 5.4 for details.

Choosing a reasonable value for the standard deviation υ_{NS} of the non-stationary regression coefficients is more difficult. For Studies 1 and 2, all of the coefficients have a true value of 0. Since few observations are used for inference, a restrictive prior should be chosen. We therefore let $\upsilon_{\mathrm{NS}} = 0.5$ in the first two studies. In Studies 3 and 4, the spatial covariates have been scaled so that the values of the non-zero regression coefficients are 1 or -1. For these studies, we use $\upsilon_{\mathrm{NS}} = 1.5$. Figure 5.6a shows the correlation structure of the process from Study 3 with $\beta_\rho = \beta_{w_x} = \beta_{w_y} = 1.5$ instead of the true values $\beta_\rho = 0$ and $\beta_{w_x} = \beta_{w_y} = 1$. Compared to Figure 5.4a, the effect of the vector field $\boldsymbol{w}(\cdot)$ is clearly stronger, and the size of the correlation structre seems to be increasing with the x-coordinate. This can also be seen in the realization shown in Figure 5.6b, where there seems the be longer dependencies in the right-hand portion of the region.

5.4 Results

The results based on 20 replication runs are presented, where new data is generated for each replication. In each study, we qualitatively compare the true correlation structure to the ones estimated during the inference. For each model, we find a replication run where the posterior estimates of the parameters are close to the average estimates, and the resulting estimated model is representative for all 20 replications. Using the posterior median of the parameters, we can then compute the correlation structure.

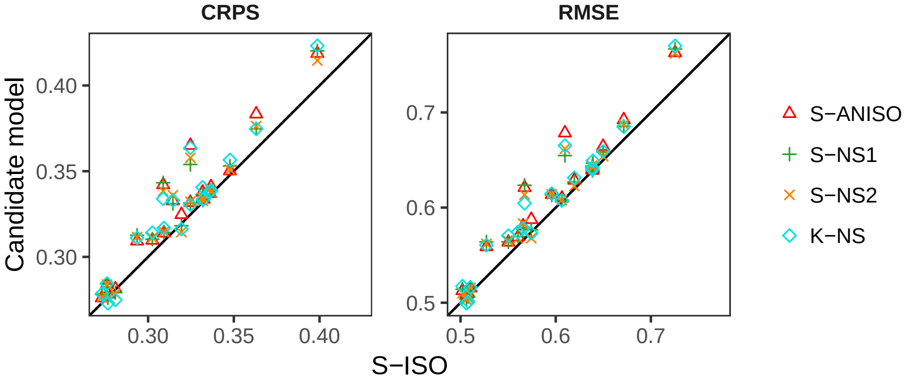


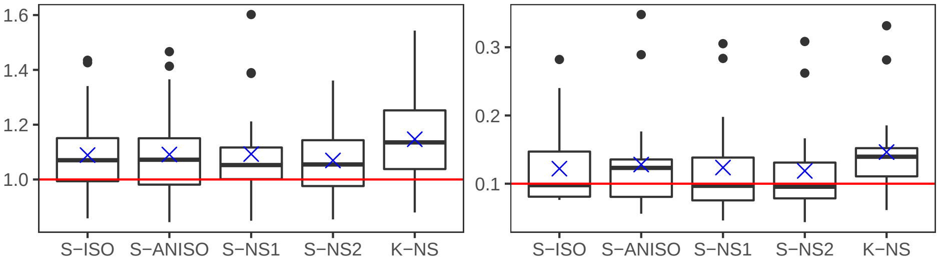
Figure 5.7: CRPS and RMSE from Study 1, using 950 prediction locations and 20 replications. Each model has one point for each replication, with the score of the candidate model along the y-axis and the score of S-ISO along the x-axis.

Table 5.2: Mean values of the prediction scores from Study 1 based on 20 replications.

Model	CRPS	RMSE
S-ISO	**0.319**	**0.588**
S-ANISO	0.329	0.605
S-NS1	0.328	0.603
S-NS2	0.328	0.602
K-NS1	0.328	0.603

Study 1 (Data from stationary, isotropic GRF)**.** In Figure 5.7 we show the CRPS and RMSE based on the 950 prediction locations for the 20 replications. Since S-ISO is equivalent to the model which $\omega(\cdot)$ was generated from, its score is plotted along the x-axis. The y-coordinate of each point is the score of the candidate model, indicated by shape and color. As a result, the points that are vertically aligned come from the same replication. The diagonal line indicates which model performed best; S-ISO is best for points above the line, while the candidate model is best for points below it.

While S-ISO almost always performs better than the more complex models, the difference is small on average. For a few select replications, some of the candidate models have marginally lower scores. The mean values of the CRPS and RMSE are listed in Table 5.2. For both scores, S-ISO leads to the lowest average, while the remaining models are slightly higher.



(a) Marginal standard deviation (b) Nugget standard deviation

Figure 5.8: Boxplots of the posterior mean estimates of the marginal and nugget standard deviation from Study 1, based on 20 replications. The mean of the estimated values (×) and the true value (—) are also shown.

Boxplots of the posterior mean estimates of the marginal standard deviation and the measurement error standard deviation σ_ε are shown in Figure 5.8, with the true value marked by a red line. The parameters are estimated well by all models, with some positive bias apparent. K-NS seems to overestimate both somewhat more than the SPDE-based models.

Figure 5.9 shows a comparison between the true correlation structure and an example of the estimated correlation structure from each model. The isotropic model S-ISO gives the correct type of correlation structure, but overestimates the effective range. In the remaining models, the baseline effective range is close to 3, which is the true value of the effective range. However, there is also clear anisotropy present, leading to a considerably longer range along one direction. The correlation structures estimated by the non-stationary models look very similar to the geometric anisotropy in S-ANISO, and there is no clear indication of non-stationarity.

△

To summarize, the best choice for prediction of an isotropic process is to use an isotropic model. However, due to a suitable choice of priors, both S-ANISO and the non-stationary models lead to predictive performances close to S-ISO. Next, we investigate if this also holds the other way around: Can an isotropic model predict data from an anisotropic model well?

Study 2 (Data from stationary, anisotropic GRF)**.** In Figure 5.10, the prediction scores after 20 replications are shown. We let the score of the true model S-ANISO be along the x-axis. One models stands out immediately, which is S-ISO. While the four other models have indistinguishable performance, S-ISO is significantly worse in every replication. This is summarized in Table 5.3, where the mean

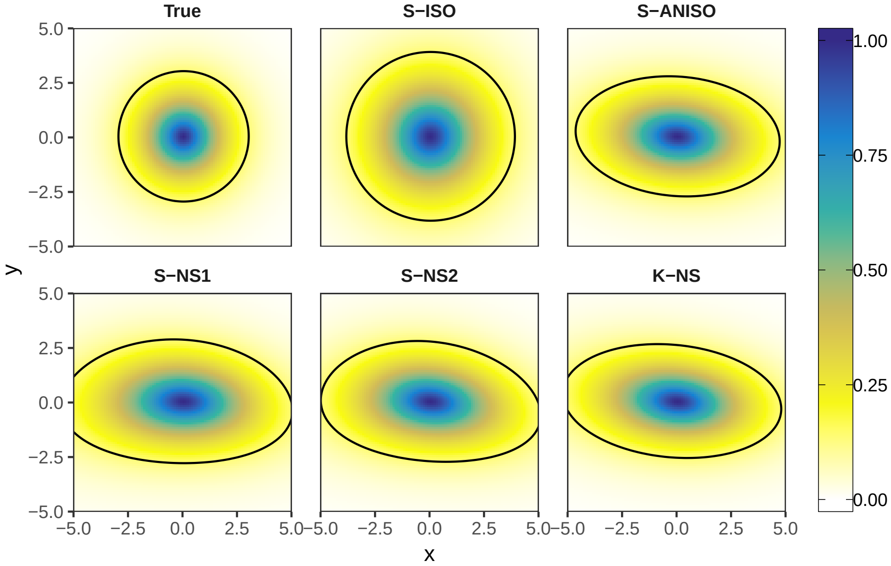


Figure 5.9: Comparison between the true and estimated correlation structures from a single replication in Study 1, with the 0.14 level curve shown in black.

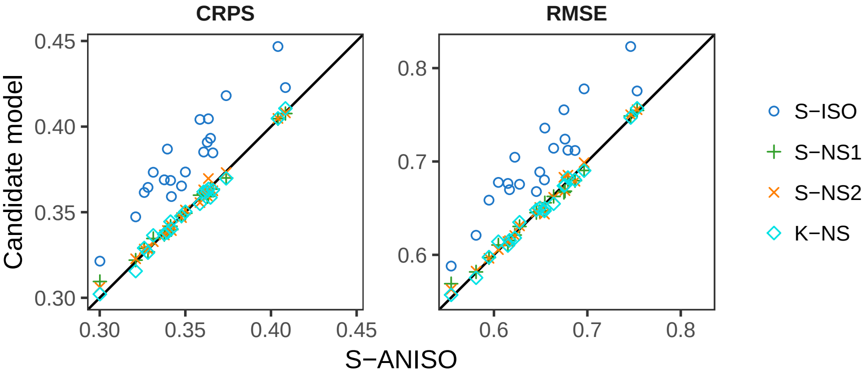


Figure 5.10: CRPS and RMSE from Study 2, using 900 prediction locations and 20 replications. Each model has one point for each replication, with the score of the candidate model along the y-axis and the score of S-ANISO along the x-axis.

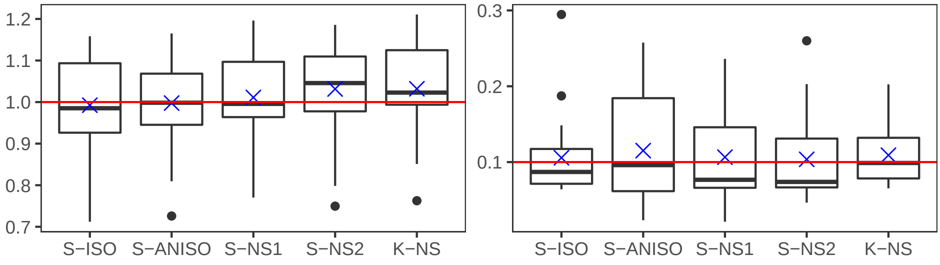
Table 5.3: Mean values of the prediction scores from Study 2 based on 20 simulation runs.

Model	CRPS	RMSE
S-ISO	<u>0.382</u>	<u>0.702</u>
S-ANISO	**0.351**	0.650
S-NS1	**0.351**	**0.649**
S-NS2	0.352	0.650
K-NS	**0.351**	0.649

CRPS and RMSE are notably higher for S-ISO, when compared to the other models.

Boxplots of both the marginal standard deviation and the nugget standard deviation are shown in Figure 5.11. These are both recovered comparably well by all five models, and the estimated bias is low.

Figure 5.12 shows a comparison between the true and estimated correlation structures. Since S-ISO is isotropic, it is not able to recover the true correlation structure. Instead, it finds a compromise by estimating an effective range somewhere between 1 and 3, the true values of the shortest and longest effective ranges. S-ANISO and the non-stationary models lead to correlation structures that are very close to the truth. While the direction of longest range is identified correctly, the strength of the anisotropy, i.e., the ratio between the longest and shortest range, is underestimated somewhat. For the non-stationary models, there is no



(a) Marginal standard deviation (b) Nugget standard deviation

Figure 5.11: Boxplots of the posterior mean estimates of the marginal and nugget standard deviation from Study 2, based on 20 replications. The mean of the estimated values (×) and the true value (—) are also shown.

indication that the estimated correlation structures are non-stationary. △

As expected, the isotropic model S-ISO leads to noticeably worse predictions when applied to anisotropic data. In addition, the non-stationary models recover the stationarity of the data, and have prediction performances comparable to the true model S-ANISO. In the next study, the direction of the longest range changes throughout the region.

Study 3 (Anisotropy with changing direction, data generated from S-NS1E)**.** The CRPS and RMSE from each replication are shown in Figure 5.13. The score of the model S-NS1 is along the x-axis. There is a clear divide between the stationary and non-stationary models. The SPDE-based models S-NS1 and S-NS2 are nearly indistinguishable, both in CPRS and RMSE. Despite having a qualitatively different correlation structure, the kernel-based K-NS is very close to S-NS1 and S-NS2 in performance. Since S-ISO and S-ANISO are unable to explain the varying direction of longest range, they lead to scores that are considerably higher. This can also be seen from Table 5.3.

The marginal standard deviation, shown in Figure 5.14a, is estimated with significantly higher bias for S-NS2, when compared to the remaining models. Figure 5.14b shows the estimated measurement error standard deviations. The stationary models seem to estimate it with somewhat higher bias that the non-stationary models.

Figure 5.15 demonstrates the true and estimated correlation structures. This is done by showing levels curves of the correlation structures centered in the grid cells closest to three points, namely $(3.25, 3.25)$, $(-3.25, 3.25)$, and $(3.25, -3.25)$. Since S-ISO and S-ANISO are stationary, the correlation structure is the same around all three points. As in Study 2, S-ISO finds a compromise by estimating a

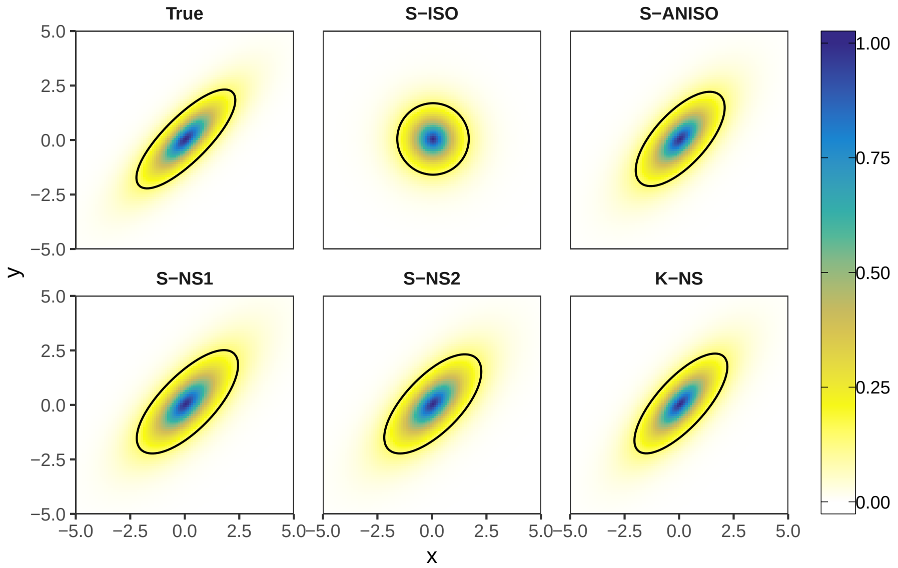


Figure 5.12: Comparison between the true and estimated correlation structures from a single replication in Study 2, with the 0.14 level curve shown in black.

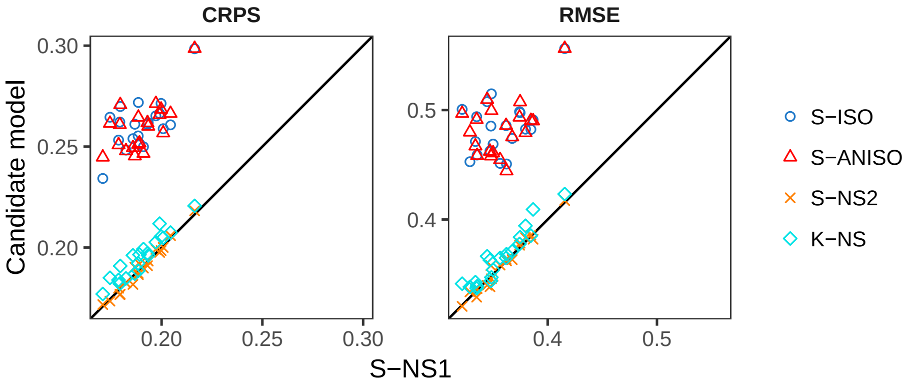
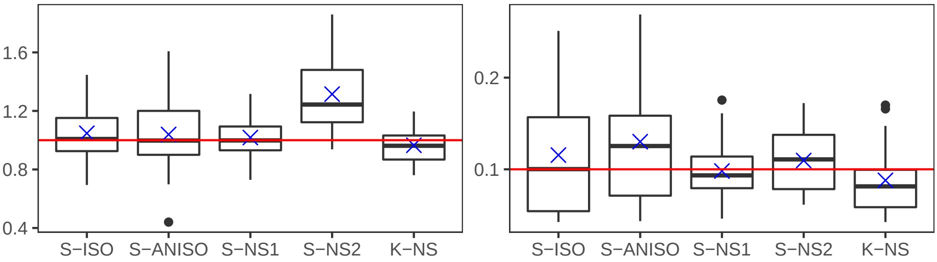


Figure 5.13: CRPS and RMSE from Study 3, using 800 prediction locations and 20 replications. Each model has one point for each replication, with the score of the candidate model along the y-axis and the score of S-NS1 along the x-axis.

Table 5.4: Mean values of the prediction scores from Study 3 based on 20 simulation runs.

Model	CRPS	RMSE
S-ISO	<u>0.261</u>	<u>0.484</u>
S-ANISO	0.260	0.483
S-NS1	**0.190**	**0.358**
S-NS2	**0.190**	**0.358**
K-NS	0.195	0.366



(a) Marginal standard deviation (b) Measurement error standard deviation

Figure 5.14: Boxplots of the posterior mean estimates of the marginal and measurement error standard deviation from Study 3, based on 20 replications. The mean of the estimated values (×) and the true value (—) are also shown.

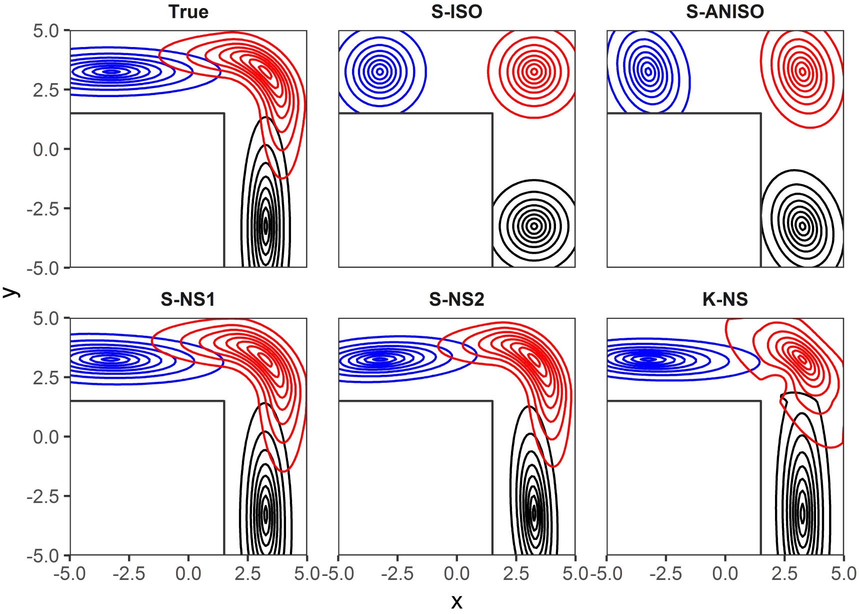


Figure 5.15: Comparison between the true and estimated correlation structures from a single replication in Study 3. The levels curves at eight levels evenly spaced between 0.14 and 0.95 are shown, centered in the grid cells closest to $(3.25, 3.25)$ (——), $(-3.25, 3.25)$ (——), and $(3.25, -3.25)$ (——).

range somewhere between the longest and shortest ranges of the true correlation structure. The same holds for S-ANISO, which also estimates some anisotropy. S-NS1 and S-NS2 both lead to correlation structures very close to the truth. The anisotropy is correctly explained by the spatial covariates given by the function $\boldsymbol{w}(\cdot)$ in Figure 5.3a, while the baseline anisotropy $(w_{x,0}, w_{y,0})$ is estimated close to 0.

K-NS leads to a qualitatively different correlation structure in the portion of the passage where the direction of $\boldsymbol{w}(\cdot)$ is changing. As a result, the correlation structure around $(3.25, 3.25)$ does not follow along the direction of $\boldsymbol{w}(\cdot)$, in contrast to S-NS1 and S-NS2. Note, however, that the correlation structure close to the center $(3.25, 3.25)$ is quite similar for all three non-stationary models. The same holds for the entire correlation structures around $(-3.25, 3.25)$ and $(3.25, -3.25)$, where $\boldsymbol{w}(\cdot)$ is constant. △

There is a clear gain in performance from using a non-stationary model when dealing with data from a non-stationary process. In addition, the kernel-based

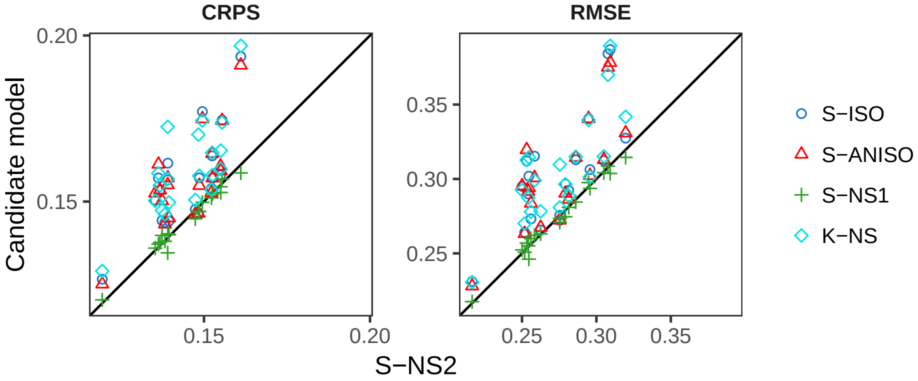


Figure 5.16: CRPS and RMSE from Study 4, using 800 prediction locations and 20 replications. Each model has one point for each replication, with the score of the candidate model along the y-axis and the score of S-NS2 along the x-axis.

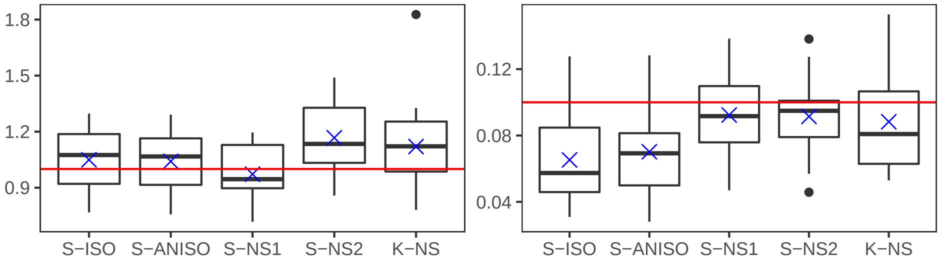
Table 5.5: Mean values of the prediction scores from Study 4 based on 20 simulation runs.

Model	CRPS	RMSE
S-ISO	0.156	0.303
S-ANISO	0.156	0.301
S-NS1	**0.145**	**0.274**
S-NS2	**0.145**	**0.274**
K-NS	0.160	0.306

model is only marginally worse than the true SPDE-based model, despite leading to qualitatively different correlation structures. The next study demonstrates a situation where the SPDE-based approach leads to better predictive performance than the kernel-based approach.

Study 4 (Barriers of short range, data generated from S-NS2)**.** In Figure 5.16 we see the CRPS and RMSE from the 20 replications. The score of the true model S-NS2 is plotted along the x-axis. Other than S-NS1, which has performance comparable to S-NS2, every model is worse. For many replications, the non-stationary K-NS is actually worse than the stationary models, despite the fact that the data comes from a non-stationary process. In addition, Table 5.5 tells us that K-NS has the highest average CRPS and RMSE.

Figure 5.17a shows the estimates of the marginal standard deviation. S-NS2



(a) Marginal standard deviation

(b) Nugget standard deviation

Figure 5.17: Boxplots of the posterior mean estimates of the marginal and nugget standard deviation from Study 4, based on 20 replications. The mean of the estimated values (×) and the true value (—) are also shown.

and K-NS lead to a somewhat higher bias than the remaining models. From Figure 5.17b we see that the stationary models struggle with the nugget standard deviation, as both models lead to an average estimate around half of the true value. The non-stationary models are, on average, much closer to the truth.

In Figure 5.18 the true and estimated correlation structures are shown. Both S-ISO and S-ANISO give reasonable estimates for the effective range, with some anisotropy apparent in S-ANISO. The correlation structure estimated with S-NS2 is close to indistinguishable from the truth. For S-NS1 the main features are recovered well, but the level curve has a different shape. K-NS seems to underestimate the effective range somewhat. It also estimates a short range along the barriers, which leads to multiple level curves.

As discussed in Example 3.1, the correlation structure obtained by K-NS is different from the one given by S-NS1 and S-NS2. With the kernel-based model, the barrier only affects the correlation close to its vicinity, which allows for dependencies through the barrier. In Figure 5.19, we consider the $20 \cdot 800 = 16000$ predictions made by each model after 20 replications. Based on these predictions, we estimate the average CRPS and RMSE as a function of distance from the nearest barrier, which is shown as a smooth line.

For all five models, the average score increases as the distance decreases. This make sense, as there are fewer observations close to the barriers. That being said, S-ISO, S-ANISO, and K-NS do significantly worse than S-NS1 and S-NS2 for small distances. The latter two are able to make the three regions independent by estimating β_ρ small enough. The former three are unable to do this, and will use observations from multiple regions when making predictions close to the barrier. Since observations from different regions are approximately independent on each other, this leads to worse predictions. △

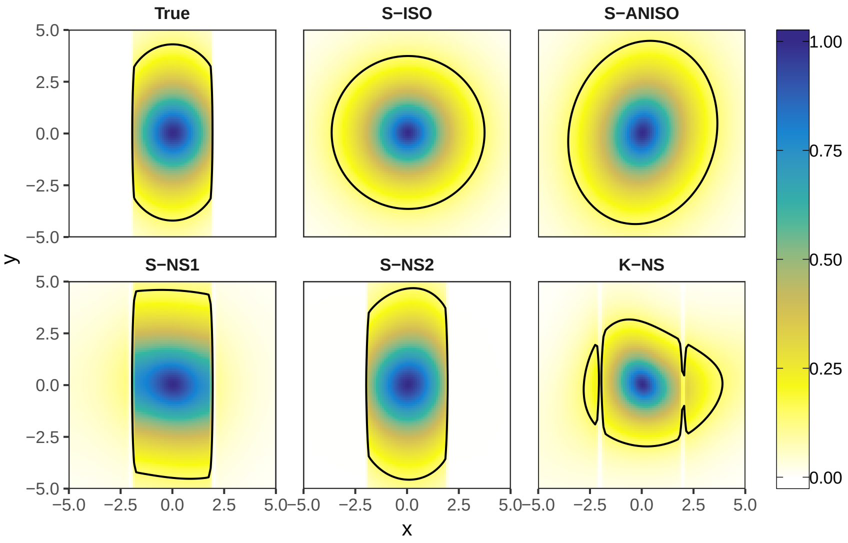


Figure 5.18: Comparison between the true and estimated correlation structures from a single replication in Study 4, with the 0.14 level curve shown in black.

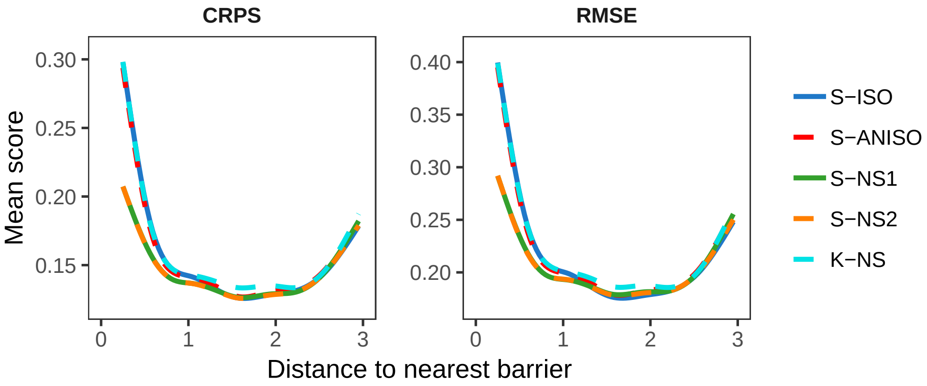


Figure 5.19: The estimated mean CRPS and RMSE from Study 4 as a function of the distance from the prediction location to the nearest barrier. For small distances, K-NS overlaps with S-ISO and S-ANISO, while S-NS1 and S-NS2 overlap at a smaller mean score.

Table 5.6: True values for parameters in Studies 1 and 3.

Study	True value						
	ρ_0	ρ_1	$w_{x,0}$	$w_{x,1}$	$w_{y,0}$	$w_{y,1}$	σ
1	3	0	0	0	0	0	1
3	1	0	0	1	0	1	1

In conclusion, using a model with the wrong type of non-stationary will lead to worse performance. It is therefore important to choose a model that can represent the type of non-stationary present in the phenomenon under consideration.

5.5 Prior sensitivity

In the parametrizations S-NS1, S-NS2, and K-NS, the functions $\rho(\cdot), w_x(\cdot)$, and $w_y(\cdot)$ are modeled by regression on spatial covariates. Choosing the standard deviation v_{NS} for the priors of the regression coefficients $\rho_1, w_{x,1}$, and $w_{y,1}$ was done in an arbitrary fashion, ensuring only that the prior covers the true value of each coefficient. The choice of v_{NS} can have significant influence on the model estimated during inference. If v_{NS} is too small, then the model might not be able to estimate the true value of the parameter. At the same time, a big value for v_{NS} can lead to a model containing superfluous parameters. This is particularly important when few observations are available, as spurious patterns in the data can be captured and explained as non-stationarity, even when the data comes from a stationary process.

In order to investigate the effect of v_{NS} on the resulting estimated model and predictions, we observe data from the stationary process in Study 1 and the non-stationary process in Study 3, and perform inference using the SPDE-based S-NS1 and the kernel-based K-NS. The covariates used for $\rho(\cdot)$, $w_x(\cdot)$, and $w_y(\cdot)$ are the same as in Study 3. The models are fitted with different values for v_{NS}, using $v_{\text{NS}} = 2^i$ for $i = -4, -3, \ldots, 3$. We observe the processes in 1000 uniformly sampled locations. First, we use use $m_{\text{inf}} = 50$ for inference and $m_{\text{pred}} = 950$ for prediction. Afterwards, we let $m_{\text{inf}} = 200$ and $m_{\text{pred}} = 800$. This is replicated 5 times, and new data is simulated for each replication.

Figure 5.20 shows the average posterior mean parameter estimates when $m_{\text{inf}} = 50$. For the parameters with multimodal posteriors ($w_{x,0}, w_{x,1}, w_{y,0}$, and $w_{y,1}$), we take the absolute value before computing the average. The non-stationary parameters are shown as dashed lines, while the remaining are shown as full lines. As v_{NS} decreases, the dashed lines all go to 0, as expected. For the stationary data, the model estimated for small values of v_{NS} is much closer to the truth than for large values. For the non-stationary data, the estimated values of $w_{x,0}$ and $w_{y,0}$ increase as $w_{x,1}$ and $w_{y,1}$ decrease, as an attempt to explain the changing direction of the anisotropy. In addition, $w_{x,1}$ and $w_{y,1}$ grow past their true value of 1 as v_{NS} increases.

In Figure 5.21, we show the CRPS based on the prediction of 950 locations. When the process is stationary, we get better predictions as v_{NS} decreases and the model is forced to become stationary. The change in CRPS, however, is very marginal. For the non-stationary process, the results are much more dramatic: forcing the model to be stationary leads to significantly worse predictions. In addition, letting v_{NS} be too large also leads to an increase in CRPS. The lowest CRPS is obtained with a v_{NS} somewhere between 0.5 and 2.

After increasing m_{inf} to 200, we obtain Figure 5.22. When the data comes from the stationary process, the effect of v_{NS} is almost non-existent. The regression coefficients are estimated close to 0 independently of the prior standard deviation. For the non-stationary data, the results are similar to what we saw with $m_{\mathrm{inf}} = 50$, only smoother. As v_{NS} decreases, both $w_{x,1}$ and $w_{y,1}$ vanish to 0, while $w_{x,0}$ and $w_{y,0}$ increase. In addition, there is little change in the estimated values for $v_{\mathrm{NS}} > 0.5$, and all of the curves flatten out.

The average CRPS for $m_{\mathrm{inf}} = 200$ is shown in Figure 5.23. The effect of v_{NS} on predictions is practically non-existent when the data is stationary, which agrees with what we observed in Figure 5.22. For the non-stationary data, the value increases sharply as v_{NS} decreases. As opposed to what we saw for $m_{\mathrm{inf}} = 50$, the curve flattens out for $v_{\mathrm{NS}} > 0.5$, which also agrees well with the average parameter estimates.

5.6 Discussion

The results from the simulation studies indicate that, for stationary data, the non-stationary models lead to prediction performance comparable to the true, stationary model. When the observed data comes from a non-stationary process, however, the non-stationary models generally lead to significantly better results. The exception is K-NS in Study 4, as the kernel-based approach leads to the wrong type of non-stationary covariance structure.

In all four studies, we generate the observed data from the SPDE-based GMRF approximation. This is also done for Studies 1 and 2, where the observed process is stationary. While it is easy to simulate from stationary GRFs, a 200×200 grid is large enough to make the differences between the GRF and GMRF negligible. A more crucial thing to note, is that the locations of the observed data coincide with centroids in the SPDE grid. When the locations are truly irregular, as is usually the case when dealing with real data, the SPDE approach must match each observation to the closest grid cell. As a result, there

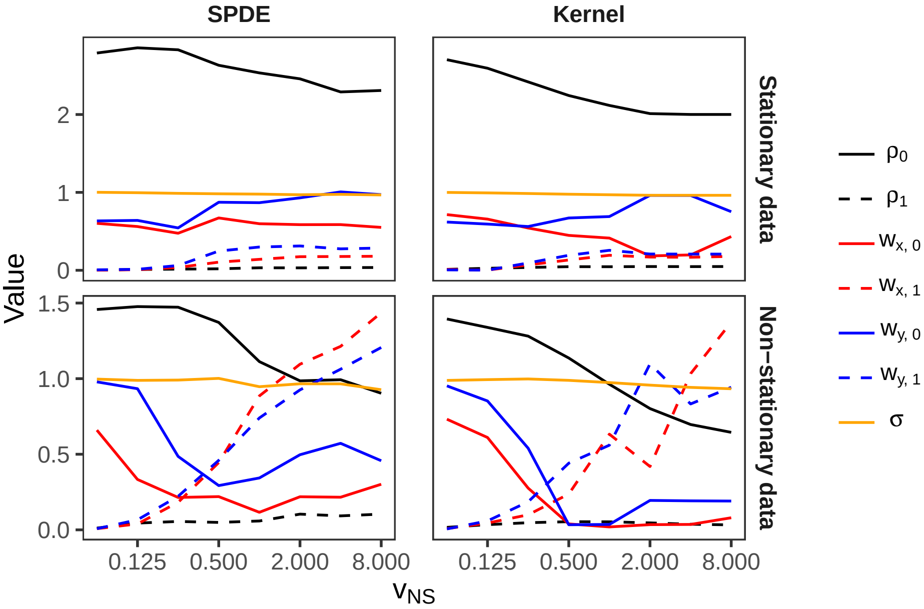


Figure 5.20: Average parameter estimates for different v_{NS}, based on 5 replications. $m_{\mathrm{inf}} = 50$ observations are used for inference. For the parameters with multimodal posteriors ($w_{x,0}, w_{x,1}, w_{y,0}$, and $w_{y,1}$), the absolute value is taken before computing the average.

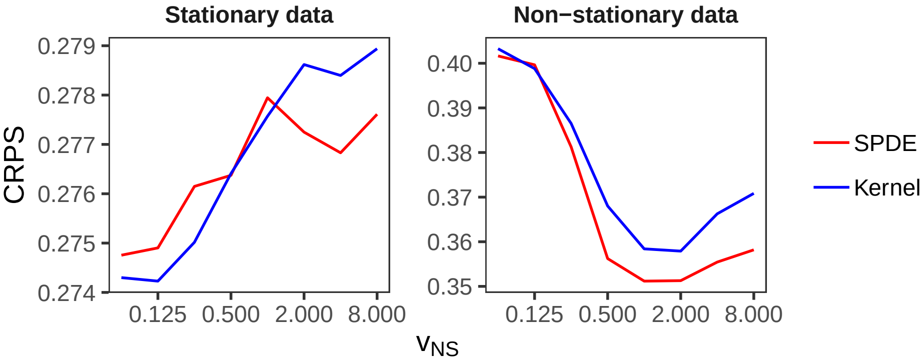


Figure 5.21: Average CRPS for different v_{NS}, based on 5 replications. $m_{\mathrm{inf}} = 50$ and $m_{\mathrm{pred}} = 950$ observations are used for inference and evaluating predictions, respectively.

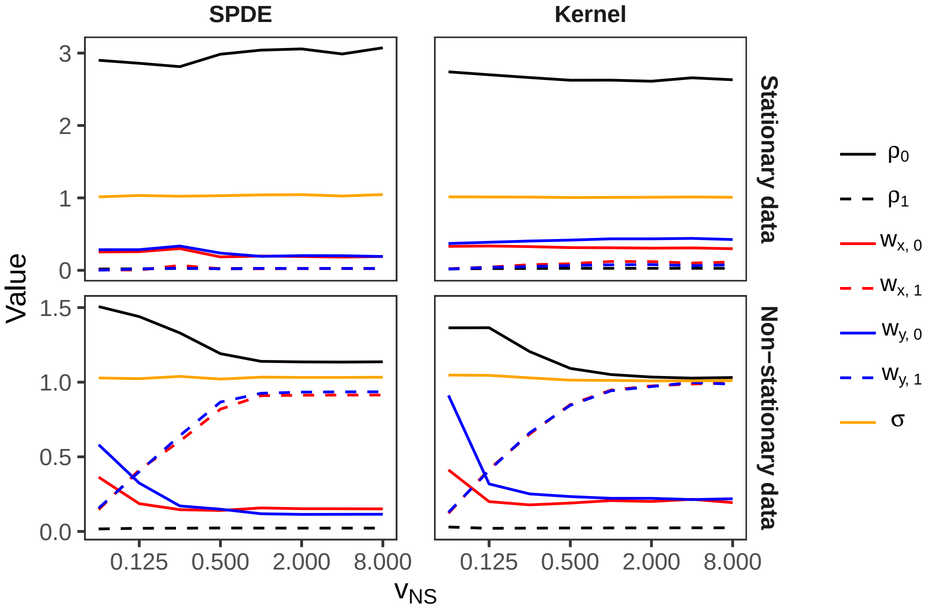


Figure 5.22: Average parameter estimates for different v_{NS}, based on 5 replications. $m_{\text{inf}} = 200$ observations are used for inference. For the parameters with multimodal posteriors, $w_{x,0}, w_{x,1}, w_{y,0}$, and $w_{y,1}$, the absolute value is taken before computing the average.

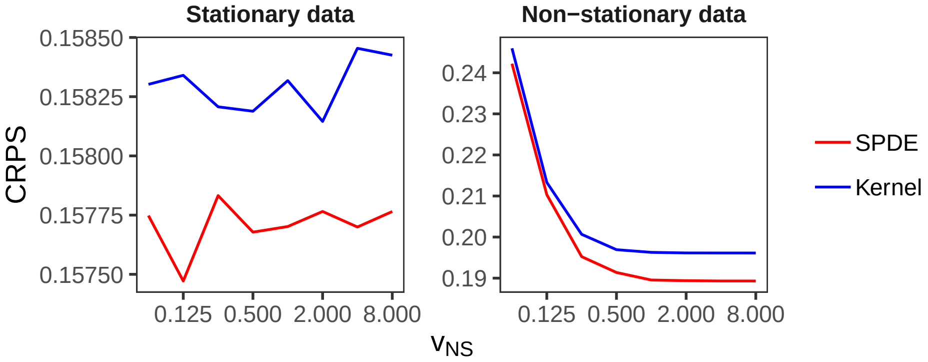


Figure 5.23: Average CRPS for different v_{NS}, based on 5 replications. $m_{\text{inf}} = 200$ and $m_{\text{pred}} = 800$ observations are used for inference and evaluating predictions, respectively.

will usually be some error due to this interpolation. The results obtained for the SPDE might therefore not reflect the performance we would get for fully irregular locations.

Separating the mean and covariance structure based on observed data is, in general, difficult. Given only a single realization, it is impossible. We focused on the role of the covariance structure by generating the observed data from processes with constant mean structure, and specifying models with mean functions consisting of only an intercept. This makes the ability to estimate the true covariance structure crucial, which was seen in Study 3. S-ISO and S-ANISO are unable to model a covariance structure with changing direction of longest range, and therefore did significantly worse than the non-stationary models in both CRPS and RMSE. In a more realistic setup, covariates would be included in the mean function.

As seen in the prior sensitivity analysis, choosing a reasonable value for the coefficient standard deviation v_{NS} is critical, as it affects both the estimated model and its predictive performance. This is particularly important when few observations are available. If v_{NS} is too small, any non-stationarity present in the data can not be captured by the model. Letting v_{NS} be too large, however, leads to non-stationary models even when the data comes from a stationary process. When more data is available, the situation changes somewhat. For stationary data, the value of v_{NS} now has more or less no effect on neither the model nor its predictions. For non-stationary data, both the parameter estimates and the predictions become worse as v_{NS} decreases towards 0. This indicates that, when the size of the observed data is large, a large value of v_{NS} should be used.

Chapter 6

Case study: Annual precipitation in the CONUS

In this chapter, we apply the parametrizations described in Chapter 3 to real data, and compare stationary and non-stationary models from the SPDE- and kernel-based approaches. This is done both by five-fold cross-validation, and by dividing the data into a grid of boxes and holding out a single box at a time. Finally, we perform inference using the full dataset, and compare both the predictions and covariance structures obtained with the best models from both approaches.

6.1 Data

We take inspiration from Risser and Turek (2019), and consider the daily average precipitation rate over the contiguous United States (CONUS) in the 2018 water year (October 1, 2017 to September 30, 2018). While the raw data is available from the Global Historical Climate Network-Daily database (Menne et al., 2012), the dataset used in Risser and Turek (2019) has not been made publicly available, and therefore had to be recreated manually. We did this downloading the daily precipitation data from 2017 and 2018, which are 1.2 GB for each year. In total, these datasets contain daily measurements from around 27000 stations in the CONUS. We extract the measurements from the CONUS and period of interest, and discard data from any measurement station with missing data in the period. That is, we only consider the stations where measurements for all 365 days available. In the end, we are left with values from 5061 measurement stations. Among these, we remove four stations that have the exact same coordinates as another station, resulting in a final number of 5057. Risser and Turek (2019), however, end up with 2311 stations in total. Closer examination indicates that none of their stations are missing from our data, and we do not investigate this discrepancy any further.

In Figure 6.1, we show the resulting daily average precipitation in millimeters per day, using a logarithmic scale on the colorbar. The elevation over the CONUS is also shown. There seems to be a clear connection between the elevation and the rate of precipitation. East of the 90°W meridian, the elevation is low, and the precipitation is very homogeneous. The values are similar in scale, and changes occur smoothly over distance. In the western region, however, there is more variation. Around the Rocky Mountains, where the elevation is high, there are both yellow and blue points clustered together, and the spatial dependencies seem to have much shorter range.

For the remainder of this chapter, we limit ourselves to the western portion of the CONUS, and remove observations east of the 90°W meridian. Our region of interest $\mathcal{D}$ is then the portion of the CONUS that lies to the west of this meridian. The resulting data consists of 3353 measurement stations, which are shown in Figure 6.2. The grey, dashed rectangle indicates the extent of the data, which contains the region of interest. In the SPDE-based models, boundary effects are avoided by using a larger rectangle. The rectangle used is shown in blue, and is obtained by extending the grey rectangle 25% in every direction.

6.2 Models

Let the data be $(y_i, \boldsymbol{s}_i)$ for $i = 1, \dots, 3353$, where y_i is the average annual precipitation and $\boldsymbol{s}_i$ is the location of the measurement station. We model this process by using the LGM described in Section 4.1, so that

$$\log(y_i) = \eta(\boldsymbol{s}_i) + \varepsilon_i, \quad i = 1, \dots, 3353.$$

The logarithm of the precipitation is used to make the Gaussian assumption more natural. The precipitation has a lower bound of 0, while its logarithm can take on values from the entire real line. The measurement errors satisfy $\varepsilon_1, \dots, \varepsilon_{3353} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$, and we let

$$\eta(\boldsymbol{s}) = \mu + \boldsymbol{x}(\boldsymbol{s})^\mathsf{T}\boldsymbol{\beta} + u(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D},$$

where μ is the intercept, $\boldsymbol{x}(\cdot)^\mathsf{T}\boldsymbol{\beta}$ is a linear effect, and $u(\cdot)$ is a spatial effect. The spatial covariates used in the linear effect are the elevation, the longitude and an interaction between these two, leading to

$$\boldsymbol{x}(\boldsymbol{s}) = \left(z_{\text{elev}}(\boldsymbol{s}), z_{\text{long}}(\boldsymbol{s}), z_{\text{elev}}(\boldsymbol{s})z_{\text{long}}(\boldsymbol{s})\right), \quad \boldsymbol{s} \in \mathcal{D},$$

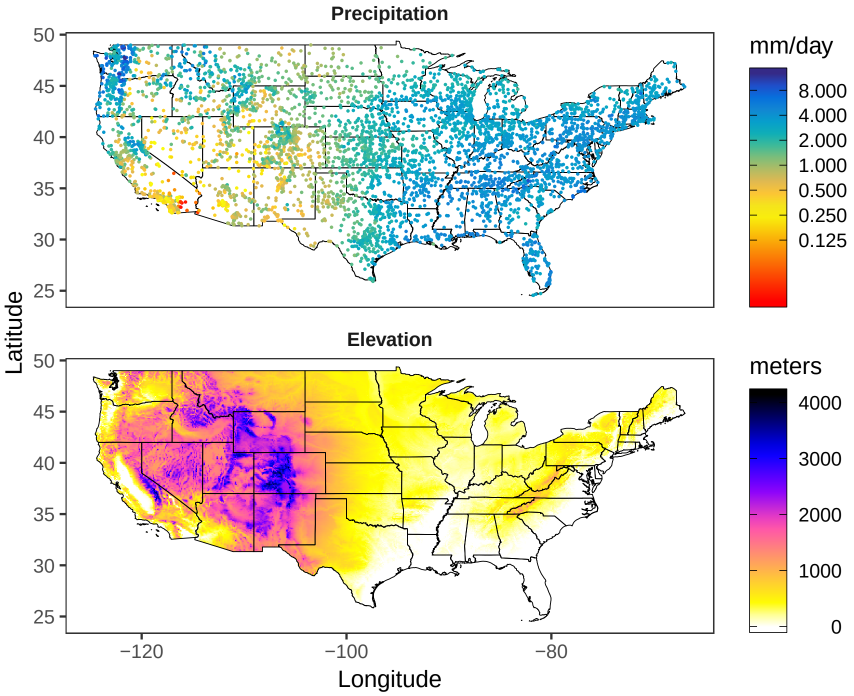


Figure 6.1: Top: Daily average precipitation in the 2018 water year, from 5057 measurements stations over the contiguous US. Note that the colors use a logarithmic scale. Bottom: Elevation over the contiguous US in meters.

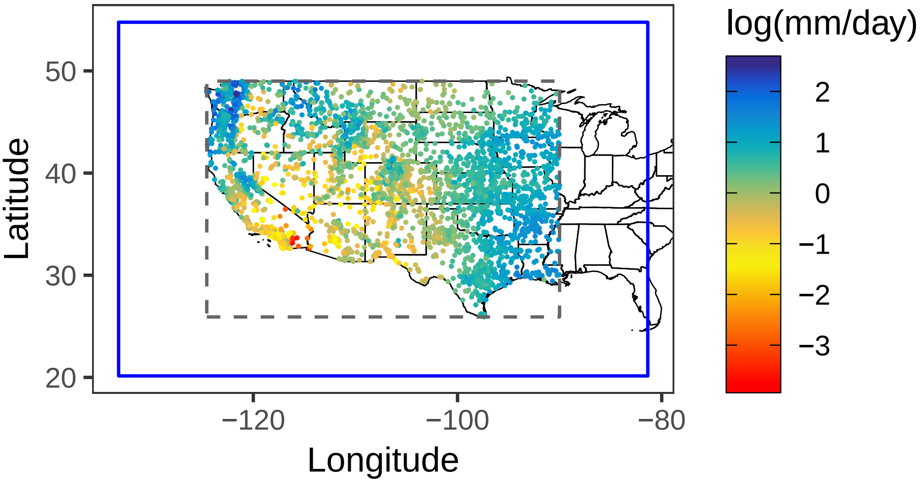


Figure 6.2: The log-daily average precipitation data from the 3353 measurement stations west of the 90°W meridian. The grey, dashed rectangle indicates the extent of the data, while the blue rectangle indicates the region used for the SPDE-based models.

where $z_{\text{elev}}(\cdot)$ and $z_{\text{long}}(\cdot)$ are functions giving the standardized elevation and standardized longitude in each location. Correspondingly, we have $\boldsymbol{\beta} = (\beta_{\text{elev}}, \beta_{\text{long}}, \beta_{\text{int}})$. A motivation for this choice can be found in Risser and Turek (2019). In essence, they argue that there seems to be a connection between precipitation and elevation, but the effect is different in, for example, Colorado and the Sierra Nevada. In the former, the areas of high elevation are dryer, while in the latter we see an increase in precipitation along the mountains. This can be adjusted for by including an interaction with the longitude. While this doesn't hold equally well for our dataset, it is nevertheless a reasonable choice.

The components discussed so far are described by 5 parameters. What remains is the spatial effect $u(\cdot)$. For this component we consider multiple candidate parametrizations, using both the SPDE- and kernel-based approaches. With the SPDE-based approach, we use a grid size of 525×350 on the blue rectangle shown in Figure 6.2. For the kernel-based approach, the Vecchia likelihood described in Section 2.3 with $k = 10$ nearest neighbors is used. A total of five models are considered for each approach. We consider the isotropic parametrizations S-ISO and K-ISO, which are described by 2 parameters, namely ρ and σ. The geometrically anisotropic S-ANISO and K-ANISO are also considered, which depend on the 4 parameters ρ, w_x, w_y, and σ. The remaining three are non-stationary, and are described in detail below.

S-VMV and K-VMV (varying marginal variance)

The non-stationary parametrizations S-NS1 and K-NS are used. In both approaches, we let the marginal standard deviation function $\sigma(\cdot)$ be modeled by the regression

$$\log(\sigma(\boldsymbol{s})) = \sigma_0 + \beta_\sigma z_{\text{elev}}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D}.$$

The effective range function $\rho(\cdot)$ and vector field $\boldsymbol{w}(\cdot) = (w_x(\cdot), w_y(\cdot))$ are both modeled as spatial constants. The covariance structure is then described by 5 parameters, namely ρ_0, $w_{x,0}$, $w_{y,0}$, σ_0, and β_σ.

S-VAN and K-VAN (varying anisotropy)

The non-stationary parametrizations S-NS1 and K-NS are used. In both approaches, we let the marginal standard deviation function $\sigma(\cdot)$ be spatially constant, while the effective range function $\rho(\cdot)$ is modeled by the regression

$$\log(\rho(\boldsymbol{s})) = \rho_0 + \beta_\rho z_{\text{elev}}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D}.$$

The vector field $\boldsymbol{w}(\cdot)$ is modeled as

$$\boldsymbol{w}(\boldsymbol{s}) = \boldsymbol{w}_0 + \beta_{\boldsymbol{w}} \boldsymbol{z}_w(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D},$$

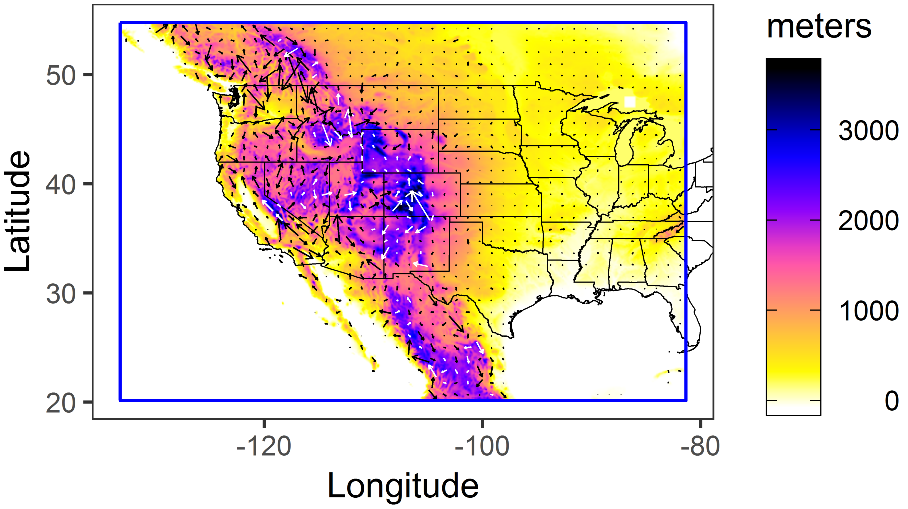


Figure 6.3: The smoothed elevation, and the vector field $\boldsymbol{z}_w(\cdot)$ scaled by a factor of 0.4, over the model region. The arrows are colored black or white, depending on the background color of the elevation.

where $\boldsymbol{w}_0 = (w_{x,0}, w_{y,0})$ and $\boldsymbol{z}_w(\cdot)$ is a vector field. For each $\boldsymbol{s}$, it is obtained by $\boldsymbol{z}_w(\boldsymbol{s}) = (-z_{y,\text{grad}}(\boldsymbol{s}), z_{x,\text{grad}}(\boldsymbol{s}))$, where $\boldsymbol{z}_{\text{grad}}(\cdot) = (z_{x,\text{grad}}(\cdot), z_{y,\text{grad}}(\cdot))$ is the gradient of a smoothed version of the elevation. In other words, $\boldsymbol{z}_w(\cdot)$ is obtained by rotating the gradient of the smoothed elevation 90 degrees counter-clockwise. The elevation is smoothed to ensure that the gradient is also reasonably smooth, and we standardize $\boldsymbol{z}_w(\cdot)$ by rescaling it so that the standard deviation of $\|\boldsymbol{z}_w(\boldsymbol{s})\|$ for $\boldsymbol{s} \in \mathcal{D}$ is 1. Both the smoothed elevation and $\boldsymbol{z}_w(\cdot)$ are shown in Figure 6.3. As we can see, the vector field tends to follow along the mountain ridges.

Note that the same coefficient $\beta_{\boldsymbol{w}}$ is used for both components of $\boldsymbol{z}_w$, resulting in a model depending on the 6 parameters ρ_0, β_ρ, $w_{x,0}$, $w_{y,0}$, $\beta_{\boldsymbol{w}}$, and σ_0.

S-FULL and K-FULL (varying anisotropy and marginal variance)

Here we use non-stationary parametrizations S-NS1 and K-NS, and combine the non-stationary models described above. The marginal standard deviation function $\sigma(\cdot)$ is modeled as in S-VMV and K-VMV, while the effective range function $\rho(\cdot)$ and vector field $\boldsymbol{w}(\cdot)$ are modeled as in S-VAN and K-VAN. This results in a model that is described by the 7 parameters ρ_0, β_ρ, $w_{x,0}$, $w_{y,0}$, $\beta_{\boldsymbol{w}}$, σ_0, and β_σ.

Table 6.1: Hyperparameters for priors of covariance structure parameters. See Section 3.3 for more details.

ρ, ρ_0		$w_x, w_y, w_{x,0}, w_{y,0}$	σ, σ_0		Regression coefficients
μ_ρ	v_ρ	$v_{\boldsymbol{w}}$	μ_σ	v_σ	v_{NS}
1.96	0.998	1.58	−0.805	0.764	3

Prior distributions

Choosing reasonable hyperparameters for the priors is more difficult here than in the simulation study in Chapter 5, as we do not know the true values of the model parameters. At the same time, the simulation study used 200 observations or less for inference, while the precipitation data consists of 3353 observations. We therefore use approach outlined in Section 3.3 and specify priors with sensible, but wide, 95% prior credible intervals for the parameters. For the effective range ρ, we use a lower bound of 1 and an upper bound of 50, so that the prior has 95% of its density between these values. The upper bound is then larger than the diagonal of the grey rectangle in Figure 6.2, which has a length of 41.6. Bounds for the marginal variance σ can be obtained by considering Figure 6.2. Based on this, we use a lower bound of 0.1 and an upper bound of 2. For the parameters w_x and w_y, we let $w_{\text{strength}} = 4$, so that $\mathrm{P}\left(\sqrt{1+w_x^2+w_y^2} > 4\right) = 0.05$. The same bounds are used for the corresponding parameters $\rho_0, w_{x,0}, w_{y,0}$, and σ_0 in the non-stationary models. We let the prior standard deviation of the non-stationary regression coefficients, v_{NS}, have a value of 3. Table 6.1 shows the resulting prior hyperparameters.

For the intercept μ and the coefficients $\boldsymbol{\beta}$ of the linear effect, we let $v_\mu = v_{\boldsymbol{\beta}} = 2$. The measurement standard deviation σ_ε uses the PC prior, and we let $U = 0.5$ and $\alpha = 0.1$ so that $\mathrm{P}(\sigma_\varepsilon > 0.5) = 0.1$.

Prediction details

There are two important details to note regarding both the posterior predictions and how these are evaluated. In the simulation study in Chapter 5, the true values of $\eta(\boldsymbol{s}_i)$ were available, and could, therefore, be used for evaluating the predictions. Now, we are forced to evaluate the predictions using the noisy observations $\log(y_i)$. Hence, instead of predicting $\eta(\boldsymbol{s}_i)$, we ensure to predict the variable $\log(y_i)$, which incorporates the uncertainty due to the measurement error ε_i. The posterior mean of the prediction is used for computing the RMSE.

When evaluating the predictions, we have two choices: we can do everything on the log-scale, i.e., compare the prediction of $\log(y_i)$ with its true value, or we can transform the prediction back to the original scale, so that we compare the prediction of y_i with its true value. The CPRS and RMSE are both in the same scale as the values used to compute the scores. We choose to evaluate the scores on the log-scale, because we want to assess the relative error in the predictions. In this way, we avoid the issues associated with the widely varying order of size that the observations have in their original scale.

6.3 Cross-validation

Cross-validation is a popular way to evaluate and compare the performance of multiple competing models. The dataset is randomly partitioned into K folds $\{A_i\}_{i=1}^K$ of roughly the same size. For each $i = 1, \ldots, K$, we leave out fold A_i and use the remaining $K - 1$ folds for model inference. Then, we use the obtained model to predict the data in fold A_i, and evaluate the predictions using the true, observed values. Based on each fold we get a score S_i, and the *cross-validation score* S is then defined as the weighted mean

$$S = \sum_{i=1}^{K} \frac{n_i}{n} S_i,$$

where $n_i = |A_i|$ is the size of fold number i and $n = \sum_{i=1}^K n_i$ is the total number of observations. It is useful to think of S as an indicator of how well the model generalizes to new and unobserved data.

In order to compare the 10 models described in Section 6.2, we perform 5-fold cross-validation. With $K = 5$, each run uses roughly 80% of the data for inference and 20% for evaluating predictions. The weighted averages of the CRPS and RMSE from the 5 runs are computed, and we obtain the cross-validation scores for each model. This is replicated 10 times in total, using a different randomly selected partition of the data each time. Based on the 10 values for each cross-validation score, we can compute an average, which estimates the mean score across all possible splits into 5 approximately equally sized folds. We also report the associated estimated standard errors of the averages. Table 6.2 shows the resulting averages and standard errors. On average, both scores see a slight improvement when using a non-stationary model. The improvement is the biggest for the models with varying anisotropy, i.e., S-VAN, S-FULL, K-VAN, and K-FULL, while S-VMV and K-VMV are only marginally better.

This is further demonstrated in Figures 6.4 and 6.4, where we show the cross-validation scores from each replication. Figure 6.4 shows only the stationary models, with the SPDE- and kernel-based models indicated using full and dashed lines, respectively. While S-ISO and S-ANISO seem to be consistently better than K-ISO and K-ANISO, the difference is very small. This indicates that,

Table 6.2: Average cross-validation scores based on 10 replications, and the standard errors of the averages. Bold indicates the lowest value for each approach, while underlined indicates the highest.

Approach	Model	CRPS (10^{-3})		RMSE (10^{-3})	
		Mean	Standard error	Mean	Standard error
SPDE	S-ISO	114.3	2.1	215.6	3.5
	S-ANISO	114.5	2.1	216.2	3.6
	S-VMV	112.5	0.9	211.9	2.0
	S-VAN	110.3	1.0	209.6	2.2
	S-FULL	**109.9**	1.0	**209.2**	2.3
Kernel	K-ISO	114.7	1.9	216.7	3.5
	K-ANISO	114.7	2.0	216.9	3.7
	K-VMV	113.5	1.0	214.4	2.7
	K-VAN	**110.2**	1.1	**211.7**	3.6
	K-FULL	111.6	1.4	216.0	4.2

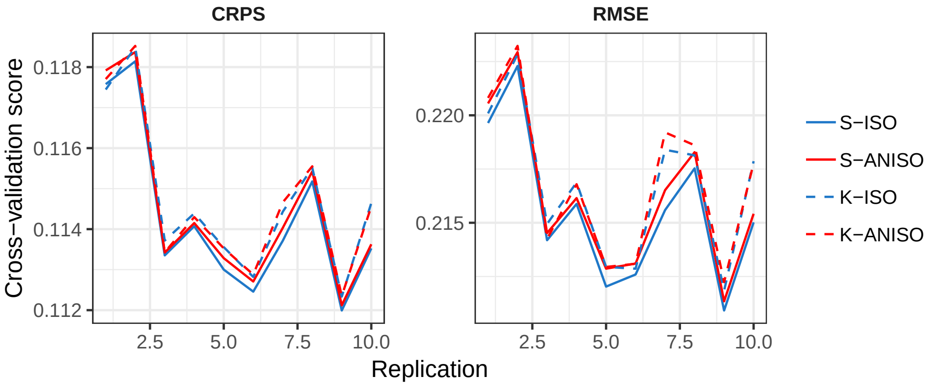


Figure 6.4: Cross-validation scores of the stationary models for each replication.

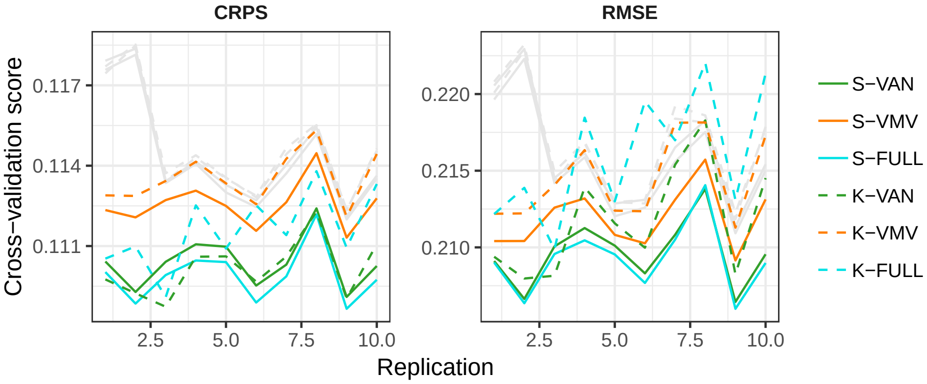


Figure 6.5: Cross-validation scores of the non-stationary models. The scores from the stationary models are shown as grey lines.

when we use models that approximate equivalent covariance structures, both approaches give comparatively good predictions. In Figure 6.5, the scores of the non-stationary models are shown, and the stationary models are included as gray lines. For the CRPS, every non-stationary model except K-VMV is consistently better than the stationary models. With the RMSE, the SPDE-based models are consistently better, while the kernel-based ones perform worse for selected replications.

Table 6.3 shows the average run-times for the different models, based on the 50 total runs. In the SPDE-based models, the run-time clearly increases with the number of parameters, and both inference and prediction takes roughly 5 times longer for S-FULL than for S-ISO. The kernel-based models take considerably longer to run, and even K-ISO is slower than S-FULL. In addition, the run-time is similar for all five models, with K-ISO being slightly faster than the rest. Naturally, this depends on a number of factors. Using a bigger grid size for the SPDEs results in longer run-times. In `BayesNSGP`, inference and prediction is done through MCMC, and the run-time grows linearly with the number of samples generated. In addition, using a larger value for the number of neighbors k also increases the run-time.

6.4 Hold-out regions

When predicting a spatial process in some new, unobserved location $\boldsymbol{s}^*$, the quality of the prediction is closely linked to the position of the new location relative to the set of observed locations. If the location is close to observed

Table 6.3: Average run times based on the 50 cross-validation runs. For the kernel-based models, we generate 50000 MCMC samples during inference, and use 5000 for prediction. All times are in seconds.

Approach	Model	#parameters	Total	Inference	Prediction
SPDE	S-ISO	7	4602	–	–
	S-ANISO	9	9071	–	–
	S-VMV	10	17144	–	–
	S-VAN	11	19488	–	–
	S-FULL	12	23575	–	–
Kernel	K-ISO	7	23578	21392	2186
	K-ANISO	9	29280	27069	2210
	K-VMV	10	30946	28726	2220
	K-VAN	11	30610	28440	2170
	K-FULL	12	31022	28834	2188

locations, the prediction will likely be good, while a more isolated location is harder to predict. When $\boldsymbol{s}^*$ is surrounded by observed data, the exact shape of the correlation structure is likely not crucial, as long as it leads to a prediction that uses the nearby observations in an effective way. The precipitation dataset contains many observations, and most are located close to other observations. It is, therefore, no surprise that all 10 models had similar performance in the cross-validation test.

In order to test how well the models predict isolated locations, i.e., locations that are not surrounded by observed data, we divide the rectangular extent of $\mathcal{D}$ into a regular 9 × 6 grid, as shown in Figure 6.6. We chose the 15 rectangles that contain the most observations, which are indicated in the figure. For each rectangle, we exclude the data inside the rectangle during inference, and use the rest as training data. Then, we predict the data in the rectangle and compare using the true, held out values.

For each of the 15 regions we obtain a CRPS and an RMSE. Based on these, we can compute the averages of the scores and their standard errors, which measures the amount of variation in the scores across the hold-out regions. This is shown in Table 6.4. As expected, the average values of the scores are higher than in Section 6.3. The variation in the scores is also considerably bigger. While every non-stationary model performs better than the stationary models on average, S-FULL and K-FULL have noticeably better scores than the best stationary models, S-ANISO and K-ANISO. Figures 6.7 and 6.8 show the scores from each run. In runs 2 and 11 the kernel-based models all have significantly higher scores, while for

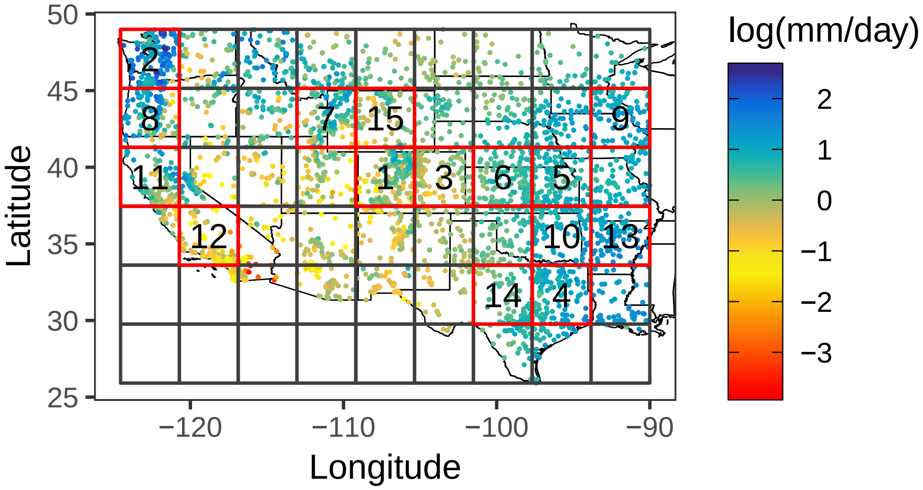


Figure 6.6: The 9×6 grid is shown in black, and the 15 rectangles containing the most observations are outlined in red. The number inside each rectangle indicates its rank by number of observations.

Table 6.4: Average hold-out scores based on 15 regions, and the standard errors of the averages. Bold indicates the lowest value for each approach, while underlined indicates the highest.

Approach	Model	CRPS (10^{-3})		RMSE (10^{-3})	
		Mean	Standard error	Mean	Standard error
SPDE	S-ISO	<u>212.4</u>	123.3	<u>356.0</u>	218.1
	S-ANISO	211.0	120.8	353.7	214.0
	S-VMV	202.0	99.79	336.6	184.3
	S-VAN	193.7	100.3	330.0	181.0
	S-FULL	**190.0**	98.07	**325.3**	177.5
Kernel	K-ISO	<u>234.8</u>	135.5	<u>394.8</u>	231.5
	K-ANISO	233.5	136.6	392.4	230.0
	K-VMV	233.2	141.7	388.6	235.5
	K-VAN	228.8	148.7	382.9	246.0
	K-FULL	**213.8**	106.0	**366.5**	187.7

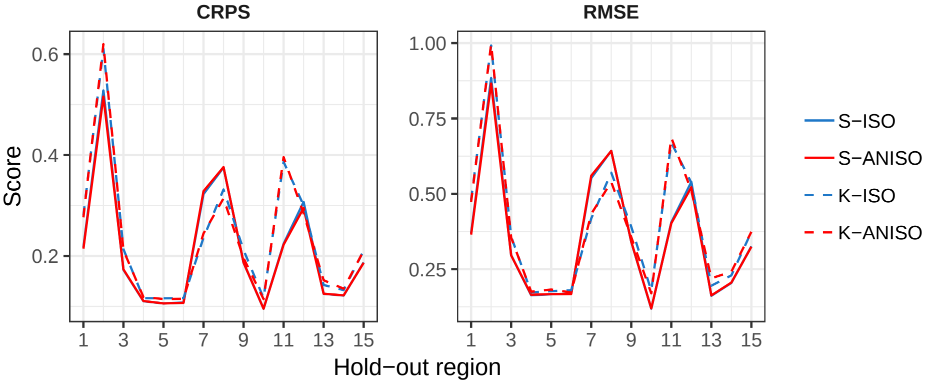


Figure 6.7: Hold-out scores of the stationary models, for the 15 regions.

runs 7 and 8 the opposite holds. This is likely due to the different approximations being used for each approach, since it holds for both the stationary and non-stationary models. The scores from the rectangles located in the eastern part of the region, such as 4, 5, 9, 10, and 13, are much lower when compared to other rectangles. This is expected, since the observations from this region are very similar in value.

6.5 Model comparison

In Sections 6.3 and 6.4 we only considered how well the different models predicted unobserved data. However, we are also interested in how the estimated models differ, especially the differences between the stationary and non-stationary models. For this purpose, we focus on a stationary (S-ANISO and K-ANISO) and a non-stationary (S-FULL and K-FULL) model from each approach. Using the entire precipitation dataset consisting of 3353 observations, we perform inference with all four models and predict the value of the log-precipitation over $\mathcal{D}$. Figure 6.9 shows the posterior mean of the predicted log-precipitation. We use the 525×350 grid over the model region in Figure 6.2 as earlier, but we only predict the value in the grid cells that are inside $\mathcal{D}$. The resulting means are indistinguishable for all four models, and they reflect the behavior of the observed data: The eastern part is relatively homogeneous and similar in value, while the middle portion has more variation. Unsurprisingly, the mean structure closely resembles the elevation from Figure 6.1, which is included as a covariate in the linear effect.

In Figure 6.10 the standard deviation of the predictions are shown. In general, the standard deviation is small near the observed locations, and large in regions

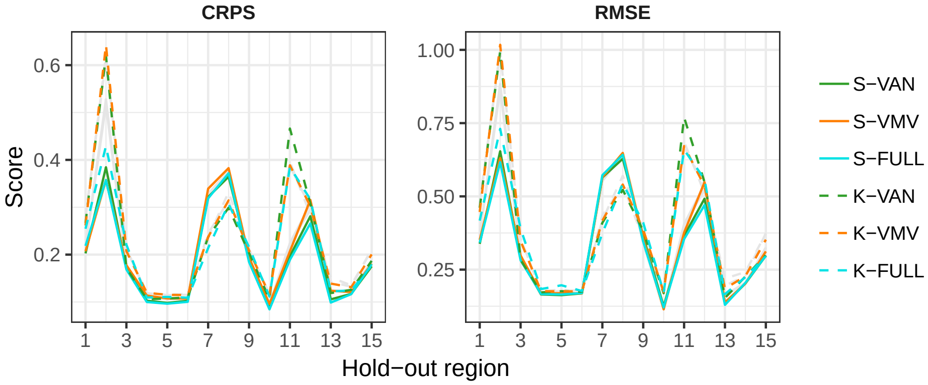


Figure 6.8: Hold-out scores of the non-stationary models, for the 15 regions. The scores from the stationary models are shown as grey lines.

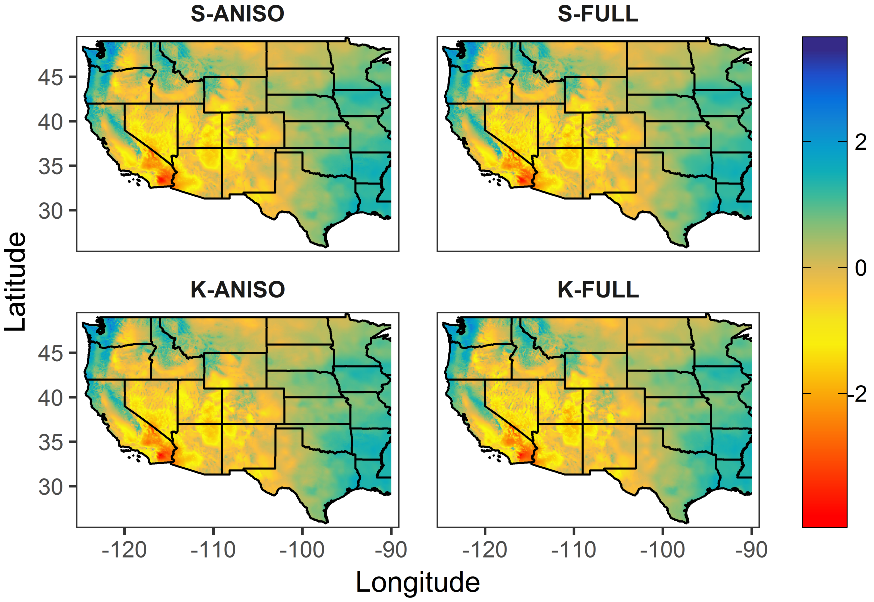


Figure 6.9: The means of the posterior predictions of the log-precipitation over $\mathcal{D}$, based on 3353 observations.

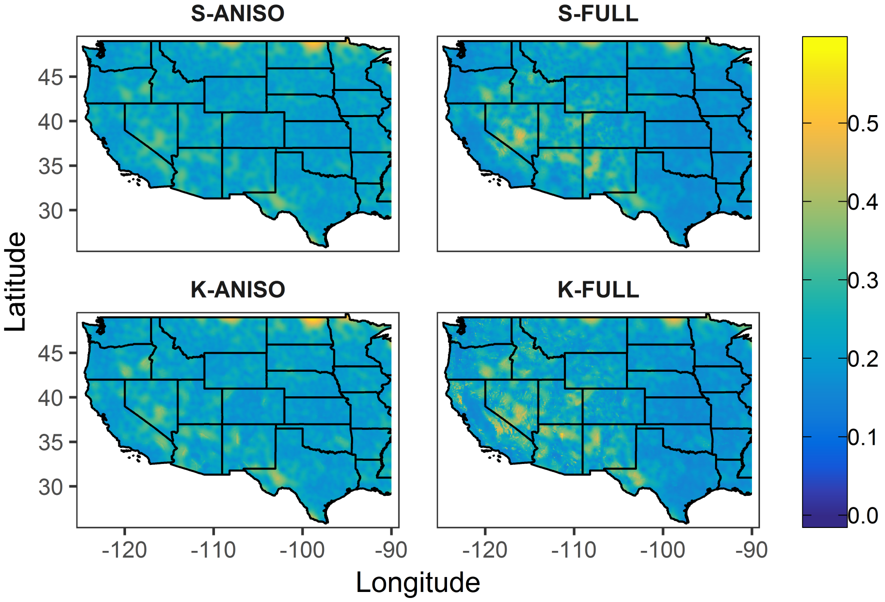


Figure 6.10: The standard deviations of the posterior predictions of the log-precipitation over $\mathcal{D}$, based on 3353 observations.

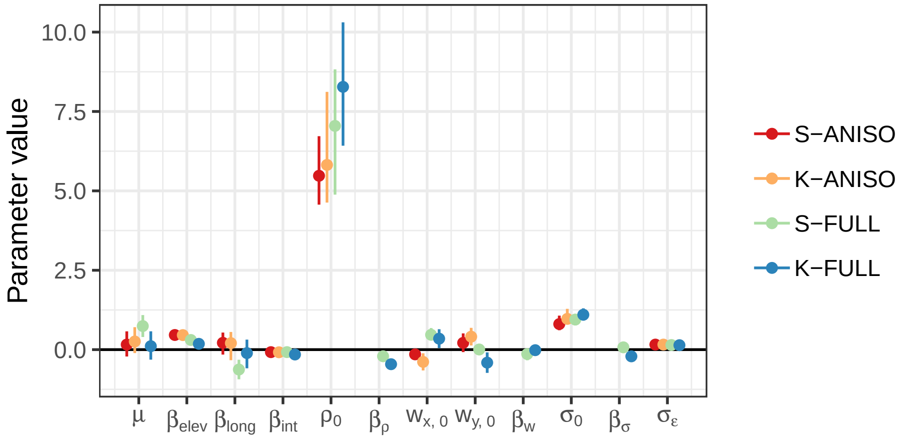


Figure 6.11: The posterior estimates of the parameters for the four models under consideration. The points indicate the median, while the lines show the $[2.5\%, 97.5\%]$ credible interval.

with no nearby observations. There is more or less no difference between S-ANISO and K-ANISO, nor between S-FULL and K-FULL. However, in some areas the non-stationary models lead to a higher standard deviation than the stationary models.

The posterior estimates of the model parameters are shown in Figure 6.11. For the most part, the models lead to similar parameter values. The marginal variance is estimated somewhere between 0.8 and 1.1, while the measurement error standard deviation is close to 0.15 for all four models. S-ANISO and K-ANISO, which approximate equivalent covariance structures, lead to more or less the same model. For S-FULL and K-FULL, the estimated β_ρ is negative, which results in a range that decreases with increasing elevation.

For this next part, we focus on the covariance structure of the component $u(\cdot)$. The correlation structures of the estimated models are demonstrated in Figure 6.12. Using the medians of the posterior parameter estimates, we compute the 0.7 isocorrelation curves centered in 18 locations throughout $\mathcal{D}$. The correlation curves of the stationary models S-ANISO and K-ANISO are very similar, with K-ANISO having a slightly longer range and more anisotropy. S-FULL has correlation curves that are close to elliptical, and the range clearly decreases with elevation. The latter also holds for K-FULL, but the shapes of the curves are far more irregular. This is investigated further in Figure 6.13, where we show the correlation structure of S-FULL and K-FULL centered in $(-110.2, 40.3)$. Despite

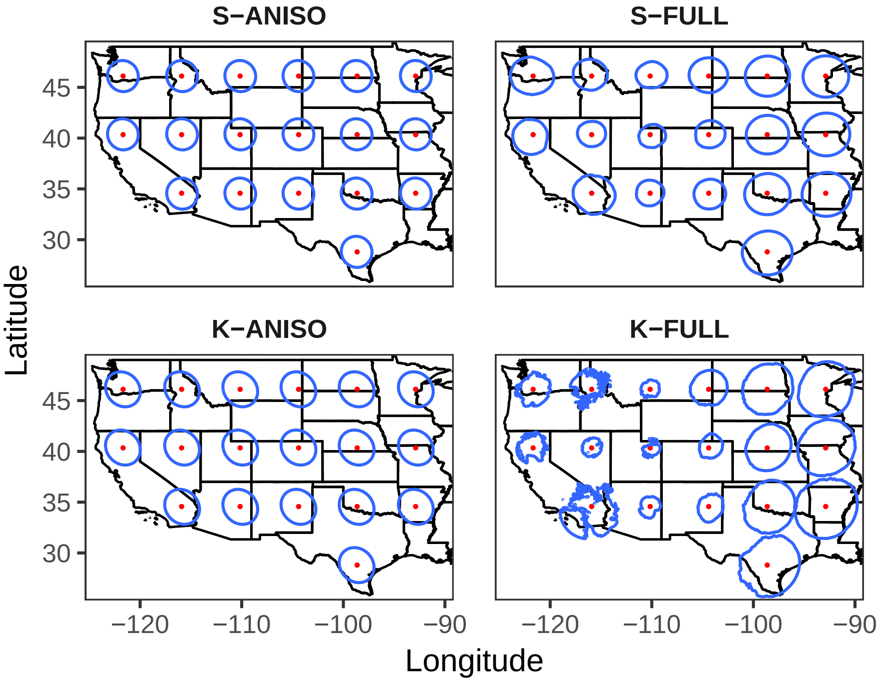


Figure 6.12: The correlation structure of $u(\cdot)$ for the four models, demonstrated by showing the 0.7 isocorrelation curves (—) centered in several locations (●). The posterior median parameter estimates are used for computing the correlations.

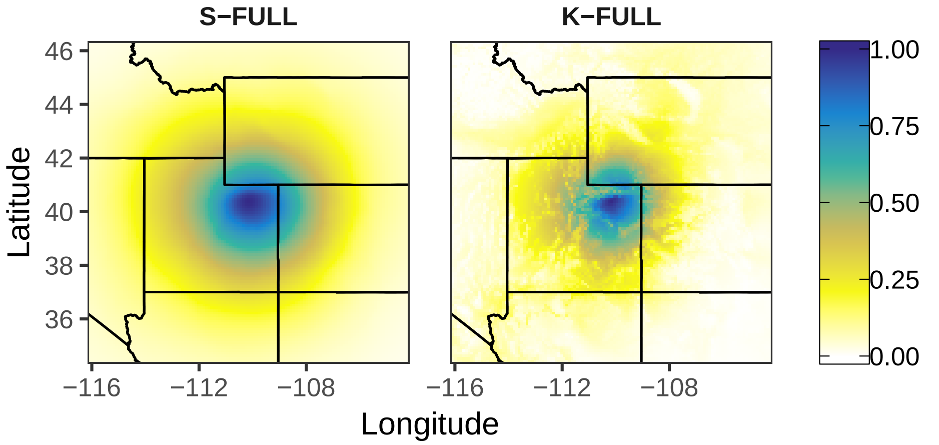


Figure 6.13: The correlation structures of $u(\cdot)$ for S-FULL and K-FULL, centered in $(-110.2, 40.3)$.

the non-smoooth elevation covariate, S-FULL leads to a smooth correlation structure that is close to isotropic. The correlation structure of K-FULL, however, clearly depends on the elevation.

While S-ANISO and K-ANISO lead to models with constant marginal variance, the same does not hold for S-FULL and K-FULL. For K-FULL, the marginal variance is given directly by the linear regression on $\sigma(\cdot)$. S-NS1, the parametrization used for S-FULL, does not control the marginal variance exactly, and the function $\sigma(\cdot)$ can only be considered an approximation. Figure 6.14 shows the exact marginal standard deviation over $\mathcal{D}$, for both S-FULL and K-FULL. The posterior median values of σ_0 and β_σ are used for the computation. In both models, the elevation is used as a covariate in $\sigma(\cdot)$. However, its effect is different: in S-FULL, the marginal variance increases with elevation, while the opposite holds for K-FULL. Also, S-FULL estimates β_σ to be smaller in absolute value, resulting in a marginal variance that varies less over $\mathcal{D}$.

6.6 Discussion

The results from the case study indicate that the non-stationary models lead to better predictions than the stationary models, both in CRPS and RMSE. In the cross-validation study in Section 6.3, the improvements were marginal. Comparing the best stationary and non-stationary model, the latter had a 4.0% (SPDE-based) and 2.7% (kernel-based) lower CRPS on average. In Section 6.4,

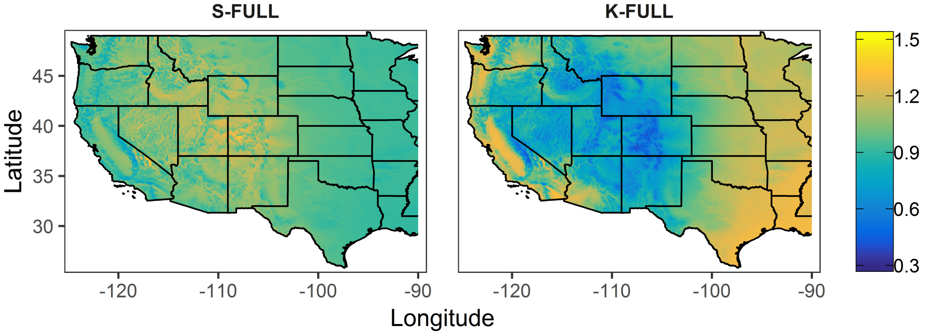


Figure 6.14: Marginal standard deviation of $u(\cdot)$ for S-FULL and K-FULL over $\mathcal{D}$, based on the the posterior medians of σ_0 and β_σ.

where entire regions of data are held out, the difference between the stationary and non-stationary models are more considerable. Here, the corresponding numbers are 8.5% and 7.2%.

In the SPDE-based models, the grid size used for the model region has a big effect on the resulting predictive performance of the model. Initially, a 300×200 grid was used. This led to the SPDE-based models performing far worse than the kernel-based models. We therefore increased the grid size until there was no more improvement, resulting in a grid size of 525×350. A similar procedure was also performed for the parameter k in the kernel-based approach, which determines the number of neighbors used in the Vecchia approximation. The value $k = 10$ was found to give a good trade-off between predictive performance and run-time.

Despite leading to comparable results, the SPDE- and kernel-based approaches have significantly different run-times. In the cross-validation study, we saw that the fastest kernel-based model was slower than the slowest SPDE-based model. The difference could, in reality, be even bigger. As mentioned, more than 50000 and 5000 samples should be used for inference and prediction with the kernel-based models. In addition, a native implementation in `R-INLA` would result in even faster run-times for the SPDE-based models.

In Section 6.5 we saw that the means of the posterior predictions were indistinguishable between both stationary and non-stationary models, and between the SPDE- and kernel-based approaches. For the prediction standard deviations, however, there are regions where the non-stationary models lead to visibly higher value. While both S-FULL and K-FULL have marginal variances for $u(\cdot)$ that clearly depend on the elevation, its effect is different. In S-FULL the marginal variance increases with the elevation, while the opposite holds for K-FULL. The effect of the spatial covariates is also different for the correlation structures. For

K-FULL, the rough nature of the elevation is clearly reflected in the irregular correlation structure. This is not the case for S-FULL, where it seems that only the size of the correlation structure is affected by the elevation.

Chapter 7

Discussion and conclusion

This thesis focuses on two approaches for specifying GRFs with spatially varying anisotropy, namely the SPDE- and kernel-based approaches. Non-stationary parametrizations based on spatial covariate regression are described for both approaches, and the resulting models are implemented in the R packages `R-INLA` and `BayesNSGP`. The qualitative differences between the approaches are demonstrated, and it is seen that the SPDE- and kernel-based approaches result in a local and global specification of the covariance structure, respectively. Inference and prediction is performed both on simulated data, and on precipitation data from the CONUS. The inferred models are compared, both in predictive performance and in estimated correlation structure, and the importance of choosing good priors is demonstrated.

Chapter 5 focuses on the simulation study, where we consider data generated from both stationary and non-stationary processes. The results from Studies 1 and 2 indicate that, given data from a stationary process, the non-stationary models from both approaches have comparable performance to true stationary model, as long as the priors are chosen carefully. In Study 3 the data is generated from a non-stationary process, and the stationary models lead to significantly worse results than the non-stationary models. While the SPDE-based S-NS1 is used for generating the data, the kernel-based K-NS is only marginally worse than S-NS1 and S-NS2 in terms of prediction performance. This is likely due to the fact that K-NS is able to recover the most crucial features of the true covariance structure, especially having a changing direction of longest range. The same does not hold for Study 4, where the true covariance structure cannot be estimated by K-NS. As a result, it has the worst predictive performance out of the 5 models that are considered. Based on this, we conclude that it is important to choose a model that is able to represent the type of non-stationarity present in the process of interest.

There are multiple aspects that have not been investigated, due to time and length constraints. In the kernel-based approach, inference and prediction is performed using MCMC. Throughout the thesis, we consistently generate 50000 samples for the inference and discard the first 5000 as burn-in. Out of the remaining 45000, we use 5000 thinned samples for prediction. Ideally, more samples should be generated. While a burn-in period of 5000 seemed sufficient for the selected MCMC chains we controlled, convergence diagnostics should be computed for all chains. Additionally, the MCMC samplers were chosen without much investigation, using Risser and Turek (2019) as a guideline. While the effect of prior width is investigated in Section 5.5, a similar analysis should also be done for Chapter 6. However, the precipitation dataset is considerably larger than the datasets used in the simulation study, and the inferred models are likely far less sensitive to prior choices.

The SPDE-based model is naturally parametrized by a sparse precision matrix, resulting in a GMRF with appealing conditional independence properties. The GMRF formulation makes the INLA framework a natural choice for performing computationally efficient inference. This is done using the `R-INLA` package, which allows for flexible specification of hierarchical Bayesian models. The kernel-based approach results in a direct construction of the covariance matrix, and computations can be made more tractable by using a nearest-neighbor Vecchia approximation. The package `BayesNSGP` is dedicated to performing inference with these kernel-based models. Here, the non-stationarity can be modeled both through regression on spatial covariates, and by representing the components of $\mathbf{H}(\cdot)$ as GRFs. While `BayesNSGP` is limited to Gaussian likelihoods, `R-INLA` has nearly 70 likelihoods available, and the SPDE model can easily by combined with other types of latent effects. In addition, `R-INLA` includes functionality for defining new latent effects and priors, which, at the moment, is not possible in `BayesNSGP`.

In this thesis, we only model non-stationarity by linear regression on spatial covariates. This assumes that there is a certain relationship between the covariance structure and the covariates, and results in an inflexible specification of both the range and the additional anisotropy. A more flexible approach is used in, for example, Fuglstad and Castruccio (2020), Fuglstad et al. (2015a), and Paciorek and Schervish (2004). In the former two, a basis function representation is utilized, while the latter models the components of the kernel function as stationary GRFs. In this way, more general covariance structures can be estimated from the observed data. However, representing the covariance structure through covariates leads to an interpretable model, and requires significantly fewer parameters than the more flexible alternatives. For example, the connection between range and elevation in Chapter 6 is intuitive, and modeling it requires only a single additional parameter.

The results from the case study in Chapter 6 indicate that non-stationary models can lead to better prediction when applied to real data. In the cross-validation study in Section 6.3, the improvements are very marginal, and the reduction in both CRPS and RMSE are at most 4.3%. In addition, there is no difference between the corresponding models from each approach, for example S-ISO and K-ISO. For the hold-out region study in Section 6.4, however, the difference between the stationary models and the best non-stationary model are considerable. There is also some difference between the predictive performance of corresponding models from the SPDE- and kernel-based approaches, as the worst SPDE-model has a lower average CRPS and RMSE than the best kernel-based model. Since S-ISO approximates the same covariance structure as K-ISO, and similarly for S-ANISO and K-ANISO, this difference is most likely due to the Vecchia approximation.

While the SPDE- and kernel-based models lead to qualitatively different covariance structures, we have no solid evidence that one approach is better than the other at prediction of real-life processes. In addition, the differences between the results obtained with the stationary and non-stationary models are small. Since the complex non-stationary models are also more time-consuming, the potential increase in predictive power might be outweighed by this increase in computation time. The SPDE-based models implemented in `R-INLA` are, nevertheless, significantly faster than the kernel-based models. The differences between the run-times become even more dramatic if a native `R-INLA` implementation is used for the SPDE models, and more MCMC samples are generated in `BayesNSGP`. Due to this, and the fact that `R-INLA` offers a wider range of functionalities than `BayesNSGP`, the SPDE approach is preferred. All in all, more investigation is necessary before any reliable conclusions can be made, and the models should be compared using other datasets than the precipitation data considered in this thesis. A particularly interesting application is non-stationary modeling of processes on the sphere, as discussed in Schmidt and Guttorp (2020). While this is described for the SPDE- and kernel-based approaches in Fuglstad and Castruccio (2020) and Heaton et al. (2014), respectively, no comparison has been made between the two resulting methods.

Bibliography

Petter Abrahamsen. A review of gaussian random fields and correlation functions, 1997.

Robert J Adler. Gaussian random fields on manifolds. In *Seminar on Stochastic Analysis, Random Fields and Applications IV*, pages 3–19. Springer, 2004.

Haakon Bakka, Jarno Vanhatalo, Janine B Illian, Daniel Simpson, and Håvard Rue. Non-stationary gaussian models with physical barriers. *Spatial statistics*, 29:268–288, 2019.

Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20, 1991.

Marta Blangiardo and Michela Cameletti. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, 2015.

Henry F Diaz, Martin P Hoerling, and Jon K Eischeid. Enso variability, teleconnections and climate change. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 21(15):1845–1862, 2001.

Geir-Arne Fuglstad and Stefano Castruccio. Compression of climate simulations with a nonstationary global spatio-temporal spde model. *The Annals of Applied Statistics*, 2020.

Geir-Arne Fuglstad, Finn Lindgren, Daniel Simpson, and Håvard Rue. Exploring a new class of non-stationary spatial gaussian random fields with varying local anisotropy. *Statistica Sinica*, pages 115–133, 2015a.

Geir-Arne Fuglstad, Daniel Simpson, Finn Lindgren, and Håvard Rue. Does non-stationary spatial data always require non-stationary random fields? *Spatial Statistics*, 14:505–531, 2015b.

Alan E Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes. *Handbook of spatial statistics.* CRC press, 2010.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Joseph Guinness. Permutation and grouping methods for sharpening gaussian process approximations. *Technometrics*, 60(4):415–429, 2018.

MJ Heaton, M Katzfuss, C Berrett, and DW Nychka. Constructing valid spatial processes on the sphere using kernel convolutions. *Environmetrics*, 25(1):2–15, 2014.

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

Peter D Hoff and Xiaoyue Niu. A covariance regression model. *Statistica Sinica*, pages 729–753, 2012.

Mathias Leander Isaksen. Fast bayesian inference for models with spatially varying anisotropy. 2019. Unpublished project report for TMA4500 Industrial Mathematics, Specialization Project.

Finn Lindgren, Håvard Rue, and Johan Lindström. 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, 2011.

Isa Marques, Nadja Klein, and Thomas Kneib. Non-stationary spatial regression for modelling monthly precipitation in germany. *Spatial Statistics*, page 100386, 2019.

Matthew J Menne, Imke Durre, Russell S Vose, Byron E Gleason, and Tamara G Houston. An overview of the global historical climatology network-daily database. *Journal of atmospheric and oceanic technology*, 29(7):897–910, 2012.

Christopher J Paciorek and Mark J Schervish. Nonstationary covariance functions for gaussian process regression. In *Advances in neural information processing systems*, pages 273–280, 2004.

Christopher J Paciorek and Mark J Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics: The official journal of the International Environmetrics Society*, 17(5):483–506, 2006.

Mark D Risser and Catherine A Calder. Regression-based covariance functions for nonstationary spatial modeling. *Environmetrics*, 26(4):284–297, 2015.

Mark D Risser and Daniel Turek. Bayesian nonstationary gaussian process modeling: the bayesnsgp package for r. *arXiv preprint arXiv:1910.14101*, 2019.

Håvard Rue, Sara Martino, and Nicolas Chopin. 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, 2009.

Håvard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005.

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

Alexandra M Schmidt and Peter Guttorp. Flexible spatial covariance functions. *Spatial Statistics*, page 100416, 2020.

Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G Martins, Sigrunn H Sørbye, et al. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, 32(1):1–28, 2017.

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

Aldo V Vecchia. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50 (2):297–312, 1988.

Peter Whittle. On stationary processes in the plane. *Biometrika*, pages 434–449, 1954.

Mathias Leander Isaksen

Comparing Global and Local Specification of Spatially Varying Anisotropy

NTNU

Norwegian University of
Science and Technology