Doctoral Thesis

Physics based dynamic modeling of space-time data

Author(s):
Sigrist, Fabio Roman Albert

Publication Date:
2013

Permanent Link:
https://doi.org/10.3929/ethz-a-009900551

Rights / License:
In Copyright - Non-Commercial Use Permitted
Acknowledgments

First of all, I would like to thank my advisors Prof. Hansruedi Künsch and Prof. Werner Stahel for giving me both excellent support and liberty when doing research. The interesting and challenging discussions we had led to many fruitful results. In a very friendly environment, Werner taught me a lot of things which include, amongst others, applied statistics. He helped me to look at things from many different perspectives and to always consider non standard approaches. I am equally thankful to Hansruedi for sharing his immense knowledge and his ability to think clearly and express things concisely.

My experience at the Seminar für Statistik (SFS) would not have been what it was without the memorable moments I had with many colleagues that have become friends by now. The good time when Jonas Peters, Patrick Mueller, Michael Chichignoud, and me shared an office passed very quickly; not just because we played a lot of backgammon and table tennis. Alain Hauser and Diego Colombo were part of the group which made the stay at the World Congress in Probability and Statistics in Istanbul in summer 2012 a very enjoyable experience. I also thank Mohammed Hebiri for playing volleyball with me, Michael Amrein and Philipp Rütimann for sharing an office, Christian Kerkhoff, Manuel Koller, and Marco Frei for enjoyable discussions at various barbecues, and Martin Mächler for excellent computer assistance. And of course, many thanks to all the people at the SFS which make the SFS the special place it is.

Furthermore, I am very grateful to Prof. Peter Guttorp and Prof. Paul Sampson for giving me the opportunity to spend four months at the University of Washington in Seattle as part of my Ph.D. studies. They made my stay in Seattle very pleasant and I have learned many things from them. One of many good memories is a joyful and hospitable Thanksgiving dinner at Paul Sampsons home. Further thanks go to my office mates in Seattle: Aaron Zimmermann for introducing me to the best Seattle coffee and the Grand Illusion cinema and Denis Leung for asking me questions I could not answer.
I also would like to thank Vanessa Stauch from Meteo Schweiz for interesting discussions during which I got to learn many things about meteorology.

In addition, I am grateful to Christian Grewe and Florian Schmidt, my two room mates during the Ph.D. studies, for good company in the spare time. Finally, I would like to say special thanks to my family and my girlfriend, Bahar Burnaz, who supported me during all the time.
Contents

Abstract ix

Zusammenfassung xi

1 Introduction 1
   1.1 Gaussian processes 2
      1.1.1 Definition and properties 3
      1.1.2 Statistical modeling 7
      1.1.3 The Gaussian assumption 8
      1.1.4 Covariance functions 9
      1.1.5 Inference and computational difficulty 12
      1.1.6 Alternatives ways of defining a Gaussian process: spectrum and stochastic differential equation 13
   1.2 Spatio-temporal models 17
      1.2.1 Realistic parametrization 18
      1.2.2 Computational tractability 19
      1.2.3 Outline 19

2 A dynamic non-stationary spatio-temporal model for short term prediction of precipitation 23
   2.1 Introduction 24
      2.1.1 Distributions for precipitation 24
      2.1.2 Correlations in space and time 25
2.1.3 Models for precipitation 25
2.1.4 Outline 26

2.2 The model 28
2.2.1 The convolution autoregressive model 29
2.2.2 Interpretation and parametrization of the kernel function 31
2.2.3 Discussion of the model 33

2.3 Fitting and prediction 35
2.3.1 Full conditional distributions 36
2.3.2 Prediction 39

2.4 Application to short term prediction of precipitation 41
2.4.1 The data 42
2.4.2 Fitting and results 44
2.4.3 Short term prediction of precipitation 46

2.5 Conclusions 52

3 SPDE based modeling of large space-time data sets 55
3.1 Introduction 56
3.2 A Continuous Space-Time Model: The Advection-Diffusion SPDE 59
3.2.1 Spectral Density and Covariance Function 62
3.2.2 Specification of the Innovation Process 63
3.2.3 Relation to an Integro-Difference Equation 64

3.3 Solution in the Spectral Space 64
3.3.1 Approximation bound 68
3.3.2 Real Fourier Functions and Discretization in Time and Space 70
3.3.3 Remarks on Finite Differences 76

3.4 Computationally Efficient Statistical Inference 77
3.4.1 Kalman Filtering and Backward Sampling in the Spectral Space 78
3.4.2 Dimension Reduction and Missing or Non-Gridded Data 81
3.4.3 An MCMC algorithm for Bayesian inference 81

3.5 Postprocessing Precipitation Forecasts 83
3.5.1 Data 85
3.5.2 Precipitation Model for Postprocessing 87
Contents

3.5.3 Fitting 89
3.5.4 Model Selection and Results 90
3.6 Conclusion 95

4 Graphical model diagnostics in Bayesian analysis using the mixed predictive distribution 97
  4.1 Introduction 97
  4.2 The mixed predictive density and its use 101
  4.3 Two simulated examples of mixed predictive distribution based graphical model checking 103
    4.3.1 A mixed effects model 103
    4.3.2 A spatio-temporal model 104
  4.4 A case study: space-time rainfall data 107

Bibliography 107

Curriculum Vitae 127
Abstract

Space-time data sets are increasingly becoming larger with more and more data being generated by radars, satellites, or highly resolved numerical models. It is therefore crucial that spatio-temporal models can deal with large data sets without running into a computational bottleneck. Statistical models in general and spatio-temporal models in particular need to have interpretable parametrizations that make a compromise between simplicity and generality. A physics based statistical approach uses knowledge from other areas of Science to build probabilistic models that meet these requirements.

In the first part of this thesis, a dynamic model for short term prediction of precipitation is presented. The model is based on a physically motivated stochastic integro-difference equation (IDE) which allows for a non-separable correlation structure that is temporally non-stationary. External meteorological variables such as wind and humidity, amongst others, are combined with a non-parametric spatio-temporal process.

Next, a model for large space-time data sets that is based on a stochastic partial differential equation (SPDE) is proposed. This approach has the advantage that the (hyper-)parameters of the process can be meaningfully interpreted, for instance, as being transport vectors and diffusion parameters. We show how to design computationally efficient algorithms by solving the SPDE in the spectral space and using the fast Fourier transform (FFT). Applying the model to postprocessing of precipitation forecasts from a numerical weather prediction (NWP)
model for northern Switzerland, we find that the statistically postprocessed forecasts outperform the raw NWP forecasts. In addition, the probabilistic forecasts appropriately quantify prediction uncertainty.

Both the IDE and the SPDE based model are formulated in a Bayesian framework and fitted using Markov chain Monte Carlo (MCMC) methods.

Furthermore, in the last part of this thesis, we demonstrate how the mixed predictive distribution (MPD) can be used for graphical model checking in Bayesian analysis. This can be applied generally for Bayesian models and, in particular, it is useful in spatio-temporal statistics.
Zusammenfassung


Danach schlagen wir ein Modell für grosse Raum-Zeit Datensätze vor, welches auf einer stochastischen partiellen Differentialgleichung beruht. Dieser Ansatz hat den Vorteil, dass die (Hyper-)Parameter des Prozesses sinnvoll interpretiert werden können, zum Beispiel als Transport-Vektor oder Diffusionsparameter. Wir zeigen, wie man re-

Beide Modelle folgen einem Bayesianischen Ansatz und werden mit Markov chain Monte Carlo Methoden an die Daten angepasst.

Chapter 1

Introduction

When modeling spatio-temporal data, there are two major challenges from a statistical, methodological point of view: first, developing realistic models that find a compromise between simplicity and generality, and second, finding algorithms that allow for applying these models in practice with a reasonable computational effort.

Concerning the first point, it is desirable to have statistical models whose parameters have a meaningful interpretation. In principle, one might presume that very general models relying on few structural assumptions can fit any data. However, such models tend to have too many (hyper-)parameters which usually cannot be identified by the data. On the other hand, some commonly used simple and convenient parametrizations are often too simple to be realistic for many real world spatio-temporal processes. Paraphrasing Einstein, models should be as simple as possible but not simpler. Other fields of science such as Physics, Ecology, Biology, Atmospheric Science, amongst others, have a lot of experience in modeling processes that exhibit variation in space and time. The idea of physics based statistical modeling is to use deterministic models as a starting point for building statistical models. Models need to be probabilistic in order to quantify uncertainty appropriately. This is of particular importance when the goal
is making predictions. The proper way of making predictions is using predictive distributions rather than point forecasts without uncertainty quantification. Uncertainty may stem from various sources such as the way things are measured, the quality of knowledge about the true underlying process, and inherent randomness of the phenomenon that is modeled.

Spatio-temporal data quickly become large as the space and / or time domain may be highly resolved. In contrast to models based on space-time covariance functions, dynamic models are Markovian in time. Apart from being physically realistic, this allows for using computationally efficient algorithms such as, for instance, the Kalman filter if the data is Gaussian. Nonetheless, if the spatial dimension is high, additional techniques are needed to cope with the computational bottleneck that makes standard computations infeasible for spatially highly resolved data.

For various reasons outlined below, Gaussian processes are used for modeling data that varies in space and time. In the following, we first give a brief introduction into Gaussian processes, followed by an overview on the particular properties and features of Gaussian processes in the spatio-temporal setting. The goal of this introduction is to make the reader familiar with the concepts on which the subsequent chapters are built. We try to keep the level of technical details rather low and, when necessary, refer to additional literature for further background.

1.1 Gaussian processes

Gaussian processes are widely used in many areas such as spatial and spatio-temporal statistics (Cressie, 1993; Gelfand et al., 2010; Cressie and Wikle, 2011) or in machine learning (Rasmussen and Williams, 2006). Conceptually, a Gaussian process can be thought of as a tool for doing non-parametric modeling, in the sense that one assumes that the function that is modeled lies in a function space rather than just being a linear combination of some covariates. As defined in more detail below, a Gaussian process is a probability distribution over an infinite dimensional function space. It is characterized by a mean and a covariance function, and samples, or realizations, from this probability
distribution are functions. In an application, the observed data and the assumptions on the particular choice of Gaussian process determine which functions are more likely and which less. In other words, with a Gaussian process one makes (prior) assumptions on the function space in which the function that is being modeled lies.

Before going into details, we start with a graphical example illustrating the use of Gaussian processes when doing non-parametric regression in a Bayesian framework. The example shows how incoming data modifies the prior function space generating a posterior function space. In Figure 1.1, the top left plot illustrates the prior Gaussian process before observing any data. The dashed lines are realizations, i.e., sample functions. The solid line shows the mean function of the prior, and the shaded area represents 95% credible intervals. The subsequent plots show how observed data determines the posterior Gaussian process. For instance, in the top right plot, 10 data points have been observed. The figure shows that close to the observation points, the uncertainty is small, and the further one moves away from data points, the larger becomes the uncertainty. As more data becomes available, the posterior uncertainty decreases, which is reflected in the gradual narrowing of the credible intervals. In this example, the data is simulated and corresponds, up to a measurement error, to a realization of the zero mean Gaussian process that is used to fit the data.

Note that we have used the Bayesian language in this example because we think it facilitates the understanding. However, Gaussian processes can be used equally well when adopting a frequentist paradigm. The process of making predictions, in general for the mean function, at a set of locations given observed data is then usually referred to as “Kriging”. For readers familiar with smoothing splines, we mention that Kimeldorf and Wahba (1971); Kent and Mardia (1994); Green and Silverman (1994); Nychka (2000) show that there are close connections between Kriging and smoothing splines.

1.1.1 Definition and properties

Defining a Gaussian process can be done in several ways with varying degrees of generality. We use the following definition from Rasmussen and Williams (2006).
Chapter 1. Introduction

Figure 1.1: Illustration of non-parametric regression using Gaussian processes. The top left plot shows the prior distribution on the function space. The following plots show how the posterior function spaces are determined by the observed data. The simulated data corresponds to a realization from a zero mean Gaussian process plus measurement error.

Definition 1 (Gaussian process). A Gaussian process is a collection of random variables \( \{ w(s); s \in \mathbb{R}^d \} \), any finite number of which have a joint Gaussian distribution.
In spatial statistics, $d$ typically equals two and Gaussian processes are often called Gaussian fields.

As an example, Figure 1.2 shows samples from two Gaussian processes. On the left, there are samples from one of the most famous Gaussian processes: the Wiener process. The right plot shows a sample from a spatial Gaussian field with a Whittle covariance function. The Wiener process is a non-stationary model (see Definition 2) with covariance function $\text{Cov}(w(s_1), w(s_2)) = \min(s_1, s_2)$, and the Whittle covariance function is a special case of the Matérn covariance function (see equation (1.5)) with smoothness parameter $\nu = 1$. See the next paragraph and Section 1.1.4 for more details on covariance functions.

![Figure 1.2: Examples of Gaussian processes: samples from a Wiener process ($d = 1$) and from a spatial Gaussian field with Whittle covariance function ($d = 2$).](image)

In the finite dimensional case, a Gaussian distribution is specified through a mean vector and a covariance matrix. Interestingly, in the infinite dimensional case, the situation is analogous: a Gaussian process $w$ is characterized by a mean function

$$m_w : \mathbb{R}^d \rightarrow \mathbb{R}$$
and a covariance function

\[ C_w : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}. \]

The mean and covariance functions are such that for all \( \{ s_1, \ldots, s_N \} \subset \mathbb{R}^d \) it holds true that

\[ (w(s_1), \ldots, w(s_N))' \sim N \left( \mu^N, \Sigma^N \right) \]

with mean vector

\[ (\mu^N)_i = \mathbb{E}[w(s_i)] = \mu_w(s_i), \quad i = 1, \ldots, N, \quad (1.1) \]

and covariance matrix

\[ (\Sigma^N)_{ij} = \text{Cov}(w(s_i), w(s_j)) = C_w(s_i, s_j), \quad i,j = 1, \ldots, N. \quad (1.2) \]

In general, in order for a function to be a covariance function it is sufficient and necessary that the function is positive definite (see Stein (1999) for more details). Further, specifying the finite dimensional distributions in (1.1) and (1.2) is indeed sufficient to characterize the infinite dimensional Gaussian process due to the Kolmogorov consistency theorem (see Gilman and Skorohod (1980); Billingsley (2012)).

The mean function specifies the deterministic mean of the process at each location \( s \), whereas the covariance function determines the dependency structure of the stochastic part of the process. Concerning the latter, heuristically speaking, the idea is that points that are geographically close together are alike. And the covariance function determines how the correlation between two points decreases with increasing distance.

In practice, one usually observes only one sample from a Gaussian process. Without any additional assumptions on the Gaussian process it is impossible to make inference from observing a single realization of a field. This is why the following two assumptions are frequently made in order to be able to estimate a model from data. Other possible assumptions to obtain an estimable model include, for instance, a decomposition as in (1.3).
1.1. Gaussian processes

Definition 2. [stationarity, isotropy] A Gaussian process is called stationary if its mean function is a constant and if its covariance function depends only on the Euclidean difference $s_1 - s_2$

$$\text{Cov}(w(s_1), w(s_2)) = C_w(s_1, s_2) = C_w(s_1 - s_2).$$

If, in addition, the covariance function only depends on the distance $|s_1 - s_2|$

$$\text{Cov}(w(s_1), w(s_2)) = C_w(s_1, s_2) = C_w(|s_1 - s_2|),$$

a Gaussian process is called isotropic.

1.1.2 Statistical modeling

In statistical applications, a Gaussian process $w(s)$ is often decomposed as

$$w(s) = x(s)^T \beta + \xi(s) + \nu(s), \quad (1.3)$$

where

- $x(s)^T \beta$ is a linear predictor ("regression term") that is related to covariates $x(s) = (x_1(s), \ldots, x_p(s))^\prime$ with coefficients $\beta \in \mathbb{R}^p$,
- $\xi(s)$ is a structured ("colored", "dependent") zero mean Gaussian process,
- $\nu(s)$ is an unstructured ("white noise", "independent") term.

This means that the mean function is modeled through a parametric regression term

$$\mu_\beta(s) = x(s)^T \beta, \quad \beta \in \mathbb{R}^p.$$ 

Variation that is not explained by this regression term is modeled non-parametrically through the zero mean Gaussian processes $\xi(s)$ and $\nu(s)$. The process $\xi(s)$ accounts for structured variation, whereas $\nu(s)$ models residual unstructured variation. The latter is thought to model measurement errors and / or small scale variability. In Geostatistics,
\( \nu(s) \) is called “nugget” effect. In practice, \( \xi(s) \) is often assumed to be a Gaussian field with a certain parametric covariance function

\[
\text{Cov}(\xi(s_1), \xi(s_2)) = C_\theta(s_1, s_2),
\]

where the parameter vector \( \theta \in \mathbb{R}^q \) determines properties of the Gaussian process. For instance, this vector may contain smoothness, marginal variance, and range parameters.

The model (1.3) can also be interpreted as a mixed effects model (Raudenbush and Bryk, 2002; Pinheiro and Bates, 2002). The parametric term \( x(s)^T \beta \) is the fixed effect, and the non-parametric Gaussian process \( \xi(s) \) corresponds to the random effect. Note that since \( \xi(s) \) is generally not observed directly, this is sometimes referred to as a latent Gaussian process.

### 1.1.3 The Gaussian assumption

In practice, the observed data need not be Gaussian. For instance, the model defined in (1.3) can be used as a component in a Bayesian hierarchical model (BHM) (Wikle et al., 1998) or in a generalized linear mixed models (GLMM). For instance, one can assume any other distribution for the observed data conditional on the linear predictor \( x(s)^T \beta + \xi(s) \). Inference becomes more challenging, though, in comparison to Gaussian data.

For instance, for modeling precipitation, one can assume the existence of a latent Gaussian process \( w(s) \), and, conditional on this, the observed precipitation \( y(s) \) is distributed according to

\[
y(s) = \max(0, w(s))^\lambda, \quad \lambda > 0.
\]

The latent variable \( w(s) \) is interpreted as a precipitation potential. If \( w(s) \) is below zero, there is no rainfall, and if \( w(s) \) is above zero, \( y(s) \) is a transformed version of \( w(s) \). In doing so, one obtains a point mass at zero modeling rainfall occurrence and a skewed distribution for the rainfall amount above zero. See Chapter 2 for more details.

The assumption that the random effect \( \xi(s) \) follows a Gaussian distribution can be justified by reasoning that there are many unmeasured
covariates whose aggregate influence is approximated by a normal distribution using an informal central limit theorem argument (Marshall and Spiegelhalter, 2007). Apart from this, the Gaussian distribution stands out from other multivariate models in the sense that it is computationally tractable in practice for reasonably large data. In addition, it is closed under marginalization which is a consistency property that is desirable for models in general.

1.1.4 Covariance functions

In order to characterize the zero mean Gaussian process $\xi(s)$ one needs to specify a covariance function. In the following, we have a closer look at covariance functions.

As said, the covariance function is usually modeled using a parametric function $C_\theta(\cdot, \cdot)$. In principle, any positive definite function can be a covariance function. In practice, it is desirable that the model be a realistic one with parameters being interpretable. Probably the most used and recommended model in spatial statistics is the Matérn covariance function (see Handcock and Stein (1993); Stein (1999)). Assuming stationarity and isotropy, this covariance function is defined as

$$C_\theta(|s_1 - s_2|) = \frac{\sigma^2}{2^{\nu-1} \Gamma(\nu)} \left( 2\sqrt{\nu} \frac{|s_1 - s_2|}{\rho} \right) \nu K_\nu \left( 2\sqrt{\nu} \frac{|s_1 - s_2|}{\rho} \right),$$

(1.5)

where $K_\nu(\cdot)$ denotes the modified Bessel function of order $\nu$.

The three parameters $\theta = (\sigma^2, \rho, \nu)'$ are interpreted as follows. The first parameter $\sigma^2$ determines the marginal variance of the process. Next, $\rho$ is a range, or scale, parameter that determines how fast the covariance decays with distance. Finally, $\nu$ is a parameter that determines the smoothness of the process.

Smoothness

To be precise, there are two kinds of smoothness for Gaussian processes: smoothness of the realizations with probability one and mean square smoothness. In general, if nothing else is said, we refer to the latter. While smoothness of the realizations implies mean square smoothness...
for Gaussian processes, the other way around this conclusion does not hold true. In order to define mean square, or \( L^2 \), smoothness, one first needs to define a corresponding limit. A sequence of random variables \( x_1, x_2, \ldots \) is said to converge in mean square, or in \( L^2 \), to a random variable \( x \) if \( \lim_{n \to \infty} \mathbb{E}[(x_n - x)^2] = 0 \) and \( \mathbb{E}[x^2] < \infty \). Consequently, a process \( w(s) \) is said to be mean square continuous if, for all points \( s \) and sequences \( s_k \) with \( \lim_{k \to \infty} s_k = s \), we have \( \lim_{k \to \infty} \mathbb{E}[(w(s) - w(s_k))^2] = 0 \). Other smoothness properties are defined analogously by looking at mean square derivatives of the process.

The relation

\[
\mathbb{E}[(w(s_1) - w(s_2))^2] = 2(C_\theta(0) - C_\theta(s_1 - s_2))
\]

shows that mean square continuity for stationary Gaussian processes is equivalent to continuity of the covariance function \( C_\theta(\cdot) \) at the origin. Other mean square smoothness properties are similarly determined by the smoothness properties of the covariance function \( C_\theta(\cdot) \) at zero. We refer to Adler (1981); Banerjee and Gelfand (2003); Banerjee et al. (2003) for more details on smoothness.

**Illustration of Matérn covariance function**

For illustration, Figure 1.3 shows how the three parameters \( \theta = (\sigma^2, \rho, \nu)' \) determine the Matérn covariance function and also corresponding samples from these processes. The three columns show the effects of varying either the variance, the range, or the smoothness parameter. In the top row, we show covariance functions with different parameters and in the bottom row samples from the corresponding Gaussian processes. For instance, looking at the right column, we see that the covariance function is smooth to a varying degree at the origin, depending on the level of \( \nu \). The corresponding samples also exhibit varying degrees of smoothness. The effect of different range parameters is illustrated in the middle column. The larger \( \rho \) the slower the covariance function decays with increasing distance. The sample paths of the processes also reflect this behaviour. The left row shows covariance functions and process realizations with three different marginal variance parameters.
Figure 1.3: Covariance functions and sample processes: effect of varying variance, range, and smoothness parameters. In the top row, Matérn covariance functions for different parameters are shown, and, in the bottom row, samples from the corresponding Gaussian processes are shown.

In principle, the covariance function $C_\xi(s_1, s_2)$ can also be modeled non-parametrically. This can be done using what is called the (semi-)variogram. See Cressie (1993); Gelfand et al. (2010) for more details. However, this approach has several disadvantages: high variability of the estimator and potential non-positive definiteness of the estimated covariance function. Furthermore, this approach cannot be used in more complex models for non-Gaussian data. In addition, the variogram does not contain all information about the local behaviour of a random field which makes it “seriously misleading for differentiable random fields” (Stein, 1999). Another approach would be to model the spectrum non-parametrically using the periodogram.
1.1.5 Inference and computational difficulty

Inference for Gaussian processes can be done in both the frequentist and the Bayesian framework. Usually, the goal is to do inference for the latent Gaussian process $\xi$ itself as well as for the regression parameters $\beta$ and the (hyper-)parameters $\theta$ that characterize the process $\xi$. Crucial for both frequentist and Bayesian inference is to evaluate the likelihood of the parameters $\theta$ and $\beta$ given observed data $w$. Concerning notation, $w$ is a vector of stacked observations where stacking is done over all points at which the discretized process used in the application is modeled. $\xi$ is defined analogously. Note that in a Bayesian hierarchical model, $w$ might not be the actually observed data, but for the full conditional of $\theta$ and $\beta$, $w$ acts as if it were the data when one uses a Markov chain Monte Carlo algorithm to sample from the posterior distribution.

The log-likelihood is then given by

$$\ell_w(\theta, \beta) = \log(P[\theta, \beta|w])$$

$$= -\frac{N}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_\theta| - \frac{1}{2} (w - \mu_\beta)\Sigma_\theta^{-1}(w - \mu_\beta),$$

(1.6)

where $N$ denotes the number of observations, and where $\mu_\beta$ and $\Sigma_\theta$ are the discretized versions of the mean and covariance function as defined in equations (1.1) and (1.2).

Equation (1.6) shows where the computational difficulty lies: calculating the determinant $\log |\Sigma_\theta|$ and the inverse $\Sigma_\theta^{-1}$. These task are best done using the Cholesky decomposition whose cost is $O(N^3)$. Starting from moderately large $N$ this task becomes computationally infeasible.

Additionally, for Bayesian inference one often needs to simulate from the full conditional distribution of the latent Gaussian process $P[\xi|\theta, \beta, w]$. One can show that, for doing so, one needs to calculate the Cholesky decomposition of the matrix

$$\left(\Sigma_\theta^{-1} + \frac{1}{\tau^2} I_N\right)^{-1}.$$
1.1. Gaussian processes

In summary, if one uses the traditional covariance based approach presented above, inference for Gaussian processes becomes computationally intractable for large data sets. The next section illustrates a way to overcome this computational bottleneck.

1.1.6 Alternatives ways of defining a Gaussian process: spectrum and stochastic differential equation

In the following, we briefly show that for defining a zero mean Gaussian process the covariance function is not the only option. This is of particular importance in spatio-temporal statistics where dynamic models play a crucial role and for large data sets where one runs into computational problems with the traditional covariance based approach. Using an example in $\mathbb{R}^1$, we illustrate how the same Gaussian process can be defined through either a covariance function, a spectral density, or a stochastic differential equation. We illustrate these concepts with a process on the real line $\mathbb{R}^1$ for the sake of simplicity. However, for higher dimensional domains $D \subseteq \mathbb{R}^d$, these approaches work analogously. In addition, one can combine these different representations, for instance, by using a spectral approach in space and a dynamic one in time to solve a stochastic partial differential equation. The advantage of having different representations for the same process (or at least approximately the same) is that one version can be computationally tractable whereas another one might not.

For the sake of understanding, we try to keep theoretical details to a minimum here. For instance, we only consider spectra which have a density with respect to the Lebesgue measure. A thorough treatment of spectral theory and linear systems requires more than just a few pages. We refer to the books of Cramér and Leadbetter (1967); Yaglom (1987) for more background. For more details on stochastic (partial) differential equations, see Øksendal (2003); Da Prato and Zabczyk (1992).

As mentioned, the following example shows that in $\mathbb{R}^1$ a stationary Gaussian process with an exponential covariance function can be defined in three equivalent ways:

- using the covariance function,
- through the spectrum,
• or as solution to a stochastic differential equation.

The nice feature of this example in \( \mathbb{R}^1 \) is that all three representations can be written explicitly in closed form. For many Gaussian processes, in particular in higher dimensions \( (d > 1) \) and in the spatio-temporal setting, these three representations exist as well but cannot always be given in closed form.

**The exponential covariance function**

The exponential covariance function is a special case of the Matérn covariance in (1.5) with smoothness parameter \( \nu = \frac{1}{2} \). In \( \mathbb{R}^1 \), the exponential covariance function can be written as

\[
Cov(\xi(s), \xi(0)) = C(s) = \pi \sigma^2 \phi^{-1} \exp(-\phi|s|), \tag{1.7}
\]

where \( \sigma^2 \) is a variance parameter and \( \phi \) a scale parameter whose reciprocal \( \rho = \phi^{-1} \) is a range parameter. In the remainder of this section, \( \xi(s) \) denotes a zero mean stationary Gaussian process with an exponential covariance function.

**The spectral density and representation**

We can calculate the Fourier transform \( (F \mathcal{C})(k) \) of the exponential covariance function to obtain what is called the spectrum, or spectral density. To be more precise,

\[
f(k) = \frac{1}{2\pi} \int_{\mathbb{R}} C(s) \exp(-iks) ds = \sigma^2 (\phi^2 + k^2)^{-1}, \tag{1.8}
\]

where \( k \in \mathbb{R} \) is a frequency (in higher dimensions, this is also called wavenumber). Applying the inverse Fourier transform, we get the covariance function back

\[
C(s) = \int_{\mathbb{R}} f(k) \exp(iks) dk. \tag{1.9}
\]

Now, using equation (1.9), we might start with any function \( f(k) \) in the first place and then obtain a covariance function. The question
is then under which conditions equation (1.9) actually gives a valid covariance function. In general, in order that $C(s)$ is a valid covariance function of a stationary, potentially complex-valued, Gaussian process, the function $f(k)$ needs to be positive and integrable. This means that for defining a zero mean stationary Gaussian process, we can use any positive and integrable function $f(k)$ which then, in turn, specifies a covariance function.

Surprisingly, this detour of first specifying a spectrum and then calculating its covariance function is actually not needed. The so called Cramér representation says that $\xi(s)$ can be represented through the stochastic integral

$$\xi(s) = \int_{\mathbb{R}} \exp(iks) d\hat{\xi}(k),$$

(1.10)

where the spectral process $\hat{\xi}(k)$ is independent white noise with variances according to (1.8)

$$\text{Cov}(\hat{\xi}(k), \hat{\xi}(l)) = \delta_{k,l} f(k).$$

Of course, the stochastic integral in (1.10) needs to be defined properly. This is done as an $L^2$ limit of sums of appropriately chosen random variables. See Cramér and Leadbetter (1967); Yaglom (1987); Billingsley (2012) for more details.

Solution to stochastic differential equation

A third representation is obtained using the following stochastic differential equation (SDE)

$$\frac{\partial}{\partial s} \xi(s) = -\phi \xi(s) + \sigma \epsilon(s),$$

(1.11)

with the stationary solution taken as initial condition, and where $\epsilon(s)$ is Gaussian white noise with unit variance. It can be shown, for instance, using the spectrum, that the solution $\xi(s)$ to this equation is a Gaussian process with an exponential covariance function.
White noise

White noise is mentioned and used above in (1.10) and (1.11). However, such a process does not exist as a regular Gaussian process. Rather it is defined as generalized process, i.e., as a random distribution or, in other words, a probability measure on a space of linear functionals. To be more specific, denote by $D(\mathbb{R})$ the space of compactly supported functions in $L^2(\mathbb{R})$. The random functional

$$D(\mathbb{R}) \rightarrow \mathbb{R}$$

$$\psi \mapsto \int \psi(s) \epsilon(s) ds$$

is white noise, if for any set $\{\psi_j | \psi_j \in D(\mathbb{R}), j = 1, \ldots, J\}$, we have

$$\left( \int \psi_1(s) \epsilon(s) ds , \ldots , \int \psi_J(s) \epsilon(s) ds \right) \sim N \left( 0, \left[ \int \psi_j(s) \psi_k(s) ds \right]_{j,k} \right).$$

Defining white noise $\epsilon(s)$ for $s \in \mathbb{R}^d$ is done analogously. For more details, see, e.g., Adler (1981); Rozanov (1982); Adler and Taylor (2007).

Discretized representations

Next, we will briefly show how the three representations presented above can be used in practice. Assume, for the sake of simplicity, that we model $\xi(s)$ at $N$ points $0, \frac{1}{N}, \ldots, \frac{N-1}{N}$ in $[0,1]$. Let $\xi$ refer to the stacked vector of the process at these points. Furthermore, denote by $\Sigma$ the finite dimensional covariance matrix of the exponential process $\xi(s)$ at these points, and let $\Sigma^{1/2}$ denote its Cholesky factorization, i.e., $\Sigma = \Sigma^{1/2} \Sigma^{T/2}$. It then follows that we can write

$$\xi = \Sigma^{1/2} \epsilon, \quad \epsilon \sim N(0, 1_N).$$

On the other hand, using the spectral representation, $\xi$ can be approximated by

$$\xi = \Phi \Sigma^{1/2} \epsilon, \quad \epsilon \sim N(0, 1_N),$$

(1.14)
where $\Phi$ denotes the discrete Fourier transform and where $\hat{\Sigma}^{1/2}$ is a diagonal matrix according to the spectrum (1.8)

$$\hat{\Sigma}^{1/2} = \text{diag}(\sqrt{f(2\pi k)}), \quad k = -(N/2 + 1), \ldots, N/2.$$ 

Furthermore, a stationary solution to the SDE in (1.11) is given by

$$\xi(i \cdot \Delta) = e^{-\Delta \phi} \xi((i-1) \cdot \Delta) + \frac{\sigma^2}{2\phi} (1-e^{-2\Delta \phi}) \epsilon(i), \quad \epsilon(i) \sim N(0, 1), \quad (1.15)$$

$i = 1, \ldots, N$, $\Delta = 1/N$, and with a suitable initial distribution for $\xi(0)$. Once can show that this representation in (1.15) gives a Cholesky decomposition of $\Sigma^{-1}$.

Now, equations (1.13), (1.14), and (1.15) each define approximately the same Gaussian process. Note that we use the word “approximately” since the processes are only equal in the limit $N \to \infty$. However, the computational costs when using these three representations in an application differ considerably. For instance, if the goal is to simulate from the distribution of the Gaussian process with exponential covariance function, the cost for the covariance based approach is $O(N^3)$ due to the Cholesky decomposition. For the spectral approach, one needs to apply the discrete Fourier transform whose cost is $O(N \log(N))$. The dynamic SDE approach is even faster with computational cost growing linearly $O(N)$.

### 1.2 Spatio-temporal models

In the following, we give a short introduction into Gaussian process in space and time. The spatial coordinates will be denoted by $s \in \mathbb{R}^d$ and the temporal one by $t \in \mathbb{R}$. The special case $d = 2$ is often used. For defining a Gaussian process in space and time, a straightforward approach would be to consider the time domain as an additional dimension and apply the Gaussian process methodology for spatial fields on $\mathbb{R}^{d+1}$ with $s' = (s, t)$. Such a covariance function based approach seems practical, since one can use all the methodology from spatial statistics.
However, there are several important arguments against this approach. Roughly speaking, the two main requirements for a spatio-temporal models are realistic parametrizations that are interpretable and computationally tractability.

As we will argue in the following, dynamic models meet both requirements, whereas covariance based models often have problems with these requirements.

1.2.1 Realistic parametrization

As said, the covariance function based approach outlined above treats time as an additional dimension in $\mathbb{R}^{d+1}$. However, this ignores important information on the etiology of the phenomenon that is modeled: time is not just another dimension. In contrast to space, time flows continuously in one direction and there is a natural order in time. In other words, time always has a physical interpretation. Ignoring this means not using valuable information. Models that explicitly consider this interpretation are dynamic ones such as the SPDE based model presented in Chapter 3. Generally, a continuous time dynamic model is a model which specifies the distribution of the instantaneous rate of change conditional on the present state. When discretizing such a model, one obtains a discrete time Markov process which defines the distribution of a future state conditional on the present and, potentially, a number of past states.

In various branches of Science, there are well known deterministic physical models that describe various processes occurring in nature. Such models include, e.g., integro-difference equations and partial differential equations. These models can be made stochastic, for instance, by adding a stochastic noise term. The corresponding stochastic models are Gaussian processes if the operators involved are linear and the additive noise is Gaussian. The advantage of such physics based models is that the parameters of the models can be given a meaningful interpretation.

Coming back to covariance functions, a simple way to construct a spatio-temporal covariance function is to take a spatial one and a temporal one and multiply them to obtain what is called a separable space-time covariance function. However, besides having unattractive
1.2. Spatio-temporal models

analytical properties (varying smoothness in different direction (Stein, 2005)), this results in a model that is too simple for many applications.

1.2.2 Computational tractability

In space and time, data sets quickly become large, either because there are many time points, many spatial points, or a combination of both. Dynamic Gaussian process models have the advantage that they are Markovian. This factorization in the likelihood allows for using computationally efficient algorithms such as the Kalman filter and the forward filtering backward sampling (FFBS) algorithm. The computational costs of these algorithms grow linear in time.

1.2.3 Outline

The remainder of this thesis is organized as follows. Chapter 2 presents a physics based dynamic model for short term prediction of precipitation. It is based on a stochastic integro-difference equation (IDE) which allows for non-separable correlation structure. At each time point, the kernel function of the integral is determined by an external wind vector. This makes the model temporally non-stationary. The underlying idea is that if the wind blows in the direction from Bern to Zurich, then rainfall in Bern should determine rainfall in Zurich at a later time stronger than the other way around. Furthermore, in the application, other meteorological covariates including temperature, dew point, and specific humidity are included in a regression term. The model allows for making probabilistic forecasts, in contrast to forecasts of numerical weather prediction models (NWP) which are deterministic.

In Chapter 3, we introduce a stochastic partial differential equation (SPDE) based model for modeling large space-time data sets. The parameters of the SPDE can be physically interpreted as explicitly modeling phenomena such as transport and diffusion that occur in many natural processes in diverse fields ranging from environmental sciences to ecology. Computational efficiency is gained by solving the SPDE using the spectral space. The fast Fourier transform allows for fast transformation between the physical and frequency space. In the spectral space, the statistical algorithms for the SPDE model are very
fast with linearly growing computational costs. We apply the model to postprocessing of precipitation forecasts from a numerical weather prediction model for northern Switzerland. The idea is that the spatio-temporal Gaussian process models the error between the true rainfall and the predicted one from the NWP in a non-parametric way. In contrast to the raw forecasts from the numerical model, the postprocessed forecasts are calibrated and quantify prediction uncertainty. Moreover, they outperform the raw forecasts. The methodology that has been developed is implemented in the R package \texttt{spate} (see Sigrist et al. (2012a)).

The IDE based model in Chapter 2 is discrete in time and continuous in space. After approximating the IDE using the Voronoi tessellation, the model becomes discrete in space as well. The SPDE based model in Chapter 3 is continuous in both space and time. As mentioned below, the models are linked in the sense that it can be shown that the SPDE model can be obtained as a limit of the IDE model when the time lag of the discrete time steps tends to zero. In Chapter 2, precipitation is modeled discretely in space at 26 locations. On the other hand, in Chapter 3, precipitation is modeled continuously in space, i.e., on a fine grid (50 × 100 quadratic grid cells with spacing of 2.2 km), which is as continuous as a digital computer can be. In Chapter 2, the goal is to predict the precipitation process by combining data from present and past precipitation observations with forecasts of other meteorological variables, which are better predictable than precipitation itself. In contrast, in Chapter 3 the aim is to postprocess precipitation forecasts. I.e., a spatio-temporal statistical model is used to quantify prediction uncertainties of deterministic precipitation forecasts from an NWP and to eliminate structured forecast errors in space and time. In both applications, the ground truth for precipitation consists of data for three-hourly rainfall amounts collected during the observation period from December 2008 to March 2009 at 26 and 32, respectively, stations in Switzerland.

When developing complex statistical models, it is important to have tools for checking these models and for guiding the search for better models. In Chapter 4, we show how the mixed predictive distribution (MPD) can be used for graphical model checking in Bayesian analy-


sis. We demonstrate how the MPD can be applied generally and, in particular, for model diagnostics in spatio-temporal statistics.

Chapters 2, 3, and 4 correspond, up to minor modifications, to the articles Sigrist et al. (2012), Sigrist et al. (2012b), and Sigrist et al. (2013).

We also mention that the work presented in Sigrist and Stahel (2011) was done in the first part of the Ph.D. studies. The use of a censored distribution for modeling Loss Given Default (LGD) data provided the initial link for the research that was done afterwards, in particular, for modeling precipitation.
Chapter 2

A dynamic non-stationary spatio-temporal model for short term prediction of precipitation

Precipitation is a complex physical process that varies in space and time. Predictions and interpolations at unobserved times and/or locations help to solve important problems in many areas. In this paper, we present a hierarchical Bayesian model for spatio-temporal data and apply it to obtain short term predictions of rainfall. The model incorporates physical knowledge about the underlying processes that determine rainfall, such as advection, diffusion, and convection. It is based on a temporal autoregressive convolution with spatially colored and temporally white innovations. By linking the advection parameter of the convolution kernel to an external wind vector, the model is temporally non-stationary. Further, it allows for non-separable and anisotropic covariance structures. With the help of the Voronoi tesselation, we construct a natural parametrization, that is space as well as time resolution consistent, for data lying on irregular grid points. In the application, the statistical model combines forecasts of three other meteorological variables obtained from a numerical weather prediction model with past precipitation observations. The model is then used
to predict three-hourly precipitation over 24 hours. It performs better than a separable, stationary, and isotropic version, and it performs comparably to a deterministic numerical weather prediction model for precipitation and has the advantage that it quantifies prediction uncertainty.

2.1 Introduction

Precipitation is a very complex phenomenon that varies in space and time, and there are many efforts to model it. Predictions and interpolations at unobserved times and/or locations obtained from such models help to solve important problems in areas such as agriculture, climate science, ecology, and hydrology. Stochastic models have the great advantage of providing not only point estimates, but also quantitative measures of uncertainty. They can be used, for instance, as stochastic generators [Wilks (1998), Makhnin and McAllister (2009)] to provide realistic inputs to flooding, runoff, and crop growth models. Moreover, they can be applied as components within general circulation models used in climate change studies [Fowler et al. (2005)], or for postprocessing precipitation forecasts [Sloughter et al. (2007)].

2.1.1 Distributions for precipitation

A characteristic feature of precipitation is that its distribution consists of a discrete component, indicating occurrence of precipitation, and a continuous one, determining the amount of precipitation. As a consequence, there are two basic statistical modeling approaches. The continuous and the discrete part are either modelled separately [Coe and Stern (1982), Wilks (1999)] or together [Bell (1987), Wilks (1990), Bardossy and Plate (1992), Hutchinson (1995), Sansó and Guenni (2004)]. Typically, in the second approach, the distribution of the rainfall amounts and the probability of rainfall are determined together using what is called a censored distribution. Originally, this idea goes back to Tobin (1958) who analyzed household expenditure on durable goods. For modeling precipitation, Stidd (1973) took up this idea and modified it by including a power-transformation for the non-zero part so that the model can account for skewness.
2.1.2 Correlations in space and time

For modeling processes that involve dependence over space and time, there are two basic approaches [see, e.g., Cressie and Wikle (2011)]: one which models the space-time covariance structure without distinguishing between the time and space dimensions, and a dynamic one which takes the natural order in the time dimension into account.

The first approach usually follows the traditional geostatistical paradigm of assuming a parametric covariance function [for an introduction into geostatistics, see, e.g., Cressie (1993) or Gelfand et al. (2010)]. Several parametric families specifying explicitly the joint space-time covariance structure have been proposed [Jones and Zhang (1997), Cressie and Huang (1999), Gneiting (2002), Ma (2003), Stein (2005), Paciorek and Schervish (2006)]. Interpretability and, especially, computational complexity are challenges when working with parametric space-time covariance functions.

There is, however, a fundamental difference between the spatial and the temporal dimensions. Whereas there is an order in the time domain, there exists no obvious order for space. It is therefore natural to assume a dynamic temporal evolution combined with a spatially correlated error term [Solna and Switzer (1996), Wikle and Cressie (1999), Huang and Hsu (2004), Xu et al. (2005), Gelfand et al. (2005)]. As Wikle and Hooten (2010) state, the dynamic approach can be used to construct realistic space-time dependency structures based on physical knowledge. Further, the temporal Markovian structure offers computational benefits.

2.1.3 Models for precipitation

Isham and Cox (1994) state that there are three broad types of mathematical models of rainfall: deterministic meteorological models [Mason (1986)], intermediate stochastic models [Le Cam (1961), Cox and Isham (1988), Waymire et al. (1984)], and empirical statistical models. Meteorological models represent as realistically as possible the physical processes involved. As noted by Kyriakidis and Journel (1999), deterministic models typically require a large number of input parameters that are difficult to determine, whereas stochastic models are usually based
on a small number of parameters. Nevertheless, statistical models can also incorporate knowledge about physical processes. Parametrizations can be chosen based on physical knowledge and covariates reflecting information about the physical processes can be included.

In the following, we briefly review statistical models for precipitation. For modeling daily precipitation at a single measuring site, Stern and Coe (1984) use a non-stationary second-order Markov chain to describe precipitation occurrence and a gamma distribution to describe rainfall amounts. Hughes and Guttorp (1994) and Hughes et al. (1999) model precipitation occurrence using a non-homogeneous hidden Markov model. With the help of an unobserved weather state they link large scale atmospheric circulation patterns with the local precipitation process. Bellone et al. (2000) and Charles et al. (1999) both extend this approach by also modeling precipitation amounts. The former propose to use gamma distributions whereas the latter use empirical distribution functions. Ailliot et al. (2009) present a hidden Markov model in combination with the transformed and censored Gaussian distribution approach used in Bardossy and Plate (1992). Also building on the same censoring idea, Sansó and Guenni (1999b) model precipitation occurrence and amount of precipitation using a transformed multivariate Gaussian model with a spatial correlation structure. Further works on statistical precipitation modeling include Sansó and Guenni (1999a), Sansó and Guenni (2000), Brown et al. (2001), Stehlik and Bardossy (2002), Allcroft and Glasbey (2003), Sloughter et al. (2007), Berrocal et al. (2008), and Fuentes et al. (2008).

2.1.4 Outline

The model presented in the following is a hierarchical Bayesian model for spatio-temporal data. At the data stage, we opt for a modeling approach that determines the discrete and the continuous parts of the precipitation distribution together. This is done by assuming the existence of a latent Gaussian variable which can be interpreted as a precipitation potential. The mean of the Gaussian variable is related to covariates through a regression term. The advantages of this one-part modeling strategy are twofold: the model contains fewer parameters and it can deal with the so called spatial (and temporal) intermittence
effect [Bardossy and Plate (1992)] which suggests smooth transitions between wet and dry areas. This means that at the edge of a dry area the amount of rainfall should be low. Wilks (1998) notes that, indeed, lower rainfall intensity is observed when more neighboring stations are dry. This feature also reflects the idea that if a model determines a low probability of rainfall for a given situation, it should also give a small expected value for its amount conditional on this event, and vice versa. However, we note that there is no consensus in the literature whether the two parts of precipitation should be modeled together or separately.

At the process level, we use a dynamic model to account for spatio-temporal variation. The model explicitly incorporates knowledge about the underlying physical processes that determine rainfall, such as advection, diffusion, and convection. Approximating an integrodifference equation, we obtain a temporally autoregressive convolution with spatially colored and temporally white innovations. The model is non-stationary, anisotropic, and it allows for non-separable covariance structures, i.e., covariance structures where spatial and temporal variation interact. While our approach builds on existing models, it includes several novel features. With the help of the Voronoi tessellation, a natural parametrization for data lying on an irregular grid is obtained. The parametrization based on this tessellation is space as well as time resolution consistent, physically realistic and allows for modeling irregularly spaced data in a natural way. To our knowledge, the use of the Voronoi tessellation for spatio-temporal data on an irregular grid is new. By linking the advection parameter of the kernel to an external wind vector, the model is temporally non-stationary.

The model is applied to predict three-hourly precipitation. The prediction model is based on three forecasted meteorological variables obtained from an NWP model as well as past rainfall observations. We compare predictions from the statistical model with the precipitation forecasts obtained from the NWP.

The remainder is organized as follows. In Section 2.2, the model specifications are presented. In Section 2.3, it is shown how the model can be fitted to data using a Markov chain Monte Carlo (MCMC) algorithm and how predictions can be obtained. Next, in Section 2.4,
Chapter 2. A dynamic non-stationary spatio-temporal model for short term prediction of precipitation

The model is applied to obtain short term predictions of three-hourly rainfall. Conclusions are given in Section 2.5.

2.2 The model

It is assumed that the rainfall \( Y_t(s) \) at time \( t \) on site \( s = (x, y)' \in \mathbb{R}^2 \) depends on a latent normal variable \( W_t(s) \) through

\[
Y_t(s) = 0, \quad \text{if } W_t(s) \leq 0, \\
= W_t(s)^\lambda, \quad \text{if } W_t(s) > 0,
\]

(2.1)

where \( \lambda > 0 \). A power transformation is needed since precipitation amounts are more skewed than a truncated normal distribution and since the scatter of the precipitation amounts increases with the average amount. The latent variable \( W_t(s) \) can be interpreted as a precipitation potential.

This latent variable \( W_t(s) \) is modeled as a Gaussian process that is specified as

\[
W_t(s) = x_t(s)^T \beta + \xi_t(s) + \nu_t(s),
\]

(2.2)

where \( \beta \in \mathbb{R}^k \), and \( \nu_t(s) \sim N(0, \tau^2) \), \( \tau^2 > 0 \), are i.i.d. The mean \( x_t(s)^T \beta \) of \( W_t(s) \) is assumed to depend linearly on regressors \( x_t(s) \in \mathbb{R}^k \). For notational convenience, we split the terms specifying the covariance function into a structured part \( \xi_t(s) \) and an unstructured “nugget” \( \nu_t(s) \). The term \( \xi_t(s) \) is a zero-mean Gaussian process that accounts for structured variation in time and space. It is specified below in Section 2.2.1. The nugget \( \nu_t(s) \) models microscale variability and measurement errors. Since, typically, the resolution of the data does not allow for distinguishing between microscale variability and measurement errors, we model these two sources of variation together. Note that the covariates \( x_t(s) \) will usually be time and location dependent. In addition to weather characteristics, Fourier harmonics can be included to account for seasonality, and functions of coordinates can account for smooth effects in space.
2.2.1 The convolution autoregressive model

We follow the dynamic approach and define an explicit time evolution through the following integrodifference equation (IDE):

\[ \xi_t(s) = \int_{\mathbb{R}^2} h_\theta(s - s') \xi_{t-1}(s') ds' + \epsilon_t(s), \quad s \in \mathbb{R}^2, \quad (2.3) \]

where \( \epsilon_t(s) \) is a Gaussian innovation that is white in time and colored in space, and \( h_\theta \) is a Gaussian kernel,

\[ h_\theta(s - s') = \phi \exp\left(- (s - s' - \mu_t)^T \Sigma^{-1} (s - s' - \mu_t) \right), \quad (2.4) \]

where the parameter vector \( \theta \) combines \( \phi \) and the elements of \( \mu_t \) and \( \Sigma^{-1} \). Note that \( \mu_t \) shifts the kernel and \( \Sigma^{-1} \) determines the range and the degree of anisotropy. The parameter \( \phi \) controls the amount of temporal correlation. More details on the interpretation of the model and specific choices of the parameters \( \mu_t \) and \( \Sigma \) are discussed below in Section 2.2.2. An illustration of this kernel can be found in the application in Section 2.4.2.

In the following, we assume that we have \( N \) measurement locations \( s_i, i = 1, \ldots, N \), where measurements are made at times \( t = 1, \ldots, T \). Instead of working with a fine spatial grid with many missing observations, we formulate an approximate model for the values at the stations only, \( \xi_t = (\xi_t(s_1), \ldots, \xi_t(s_N))^T \). Discretizing the integral in (3.7), we obtain

\[ \int_{\mathbb{R}^2} h_\theta(s_i - s') \xi_{t-1}(s') ds' \approx \int_A h_\theta(s_i - s') \xi_{t-1}(s') ds' \approx \sum_{j=1}^N h_\theta(s_i - s_j) \xi_{t-1}(s_j)|A_j|. \quad (2.5) \]

Here \( A \subset \mathbb{R}^2 \) is an area which contains the convex hull of all stations, the sets \( A_i, i = 1, \ldots, N \) form a tessellation of \( A \) with \( s_i \in A_i \) and \( |A_j| \) denotes the area of cell \( A_j \).

Our model can then be written as the vector autoregression

\[ \xi_t = \phi G_t \xi_{t-1} + \epsilon_t, \quad G_t \in \mathbb{R}^{N \times N}, \quad (2.6) \]
Chapter 2. A dynamic non-stationary spatio-temporal model for short term prediction of precipitation

where

\[(G_t)_{ij} = \exp \left( -(s_i - s_j' - \mu_t)^T \Sigma^{-1} (s_i - s_j' - \mu_t) \right) \cdot |A_j|, \quad (2.7)\]

and where \(\epsilon_t = (\epsilon_t(s_1), \ldots, \epsilon_t(s_N))'\).

Note that this process does not exhibit explosive growth if the largest eigenvalue of \(\phi G_t\) is smaller than one. To ensure this, we check in our application that the largest eigenvalue is smaller than one for the parameters at the posterior modes.

**Figure 2.1:** Locations of stations. Both axes are in km using the Swiss coordinate system (CH1903). The lines illustrate the Voronoi tessellation. Cells with unbounded area have been replaced by circles whose area is determined as described in the text.

If the \(s_i\)'s form a regular grid, a tessellation is straightforward. Otherwise, we propose to use the Voronoi tessellation [Voronoï (1908)]
2.2. The model

which decomposes the space. Specifically, each site \( s_i \) has a corresponding Voronoi cell consisting of all points closer to \( s_i \) than to any other site \( s_j, j \neq i \) [see, e.g., Okabe et al. (2000) for more details]. Stations on the boundary of the convex hull have cells with infinite area. For these stations, we define \( |A_i| \) as described in the following. We first calculate the Voronoi tessellation of \( \mathbb{R}^2 \). We then replace unbounded cells by cells whose area is the average area of the neighbouring bounded cells. In Figure 3.5, the Voronoi tessellation for the Swiss stations used in the application below is shown as an example. As said, concerning the stations on the boundary, the unbounded cells have been replaced by cells \( A_i \) whose area is determined as described above. The circles represent the surface area \( |A_i| \).

As mentioned before, the \( \epsilon_t \)'s are assumed to be independent over time and colored in space. More precisely, we assume a stationary, isotropic Gaussian random field

\[
\epsilon_t \sim N(0, \sigma^2 V_{\rho_0}), \quad \sigma^2 > 0,
\]

with

\[
(V_{\rho_0})_{ij} = \exp\left(-d_{ij}/\rho_0\right), \quad \rho_0 > 0, \quad 1 \leq i, j \leq N,
\]

where \( d_{ij} \) denotes the Euclidean distance between two sites \( i \) and \( j \). The exponential correlation function is used for computational convenience. In principle, it is possible to use other covariance functions, for instance, other members of the Matérn family.

The approximation in (2.5) assumes that \( h_{\theta} \) is approximately constant in each cell. If some cells are considered to be too large for this approximation to be reasonable, additional points \( s^*_j \) can be added for which all observations are missing. Since such additional points increase the computational load, some compromise has to be found between accuracy and computational feasibility.

2.2.2 Interpretation and parametrization of the kernel function

For the purpose of interpretation, we note that, in the limit when the temporal spacing goes to zero, the solution of the IDE (3.7) can also
be written as the solution of the stochastic partial differential equation (SPDE) (see Brown et al. (2000))

$$
\frac{\partial}{\partial t}\xi_t(s) = -\mu_t \cdot \nabla \xi_t(s) + \frac{1}{4} \nabla \cdot \Sigma \nabla \xi_t(s) - \eta \xi_t(s) + B_t(s),
$$

(2.10)

where $\nabla = \left( \frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right)$ is the gradient operator and where $B_t(s)$ is temporally independent and spatially dependent. The terms have the following interpretations: $\mu_t \cdot \nabla \xi_t(s)$ models advection, $\mu_t$ being a drift or velocity vector. The second term is a diffusion term that can incorporate anisotropy, and $-\eta \xi_t(s)$ accounts for damping. The damping parameter $\eta$ is related to $\phi$ and $\Sigma$ through $\eta = -\log(\phi \pi |\Sigma|^{1/2})$. $B_t(s)$ is a source-sink or stochastic forcing term that can be interpreted as modeling convective phenomena. This interpretation is based on the reasoning that typically convective precipitation cells emerge and cease on the domain of interest in contrast to larger scale advective precipitation that is being transported over the area.

We now turn to the discussion of the parameterization of $\mu_t$ and $\Sigma$. In our application, we have information about wind. It is assumed that the drift term $\mu_t$ is proportional to this external wind vector. With $\mu_t$ varying over time, the model is temporally non-stationary. It is also conceivable that in certain situations $\Sigma$ or $\eta$ may vary over time and/or space, thus obtaining different forms of non-stationarity. Concerning $\Sigma$, it is thought that potential anisotropy is related to topography. Denoting by $w_t$ the wind vector at time $t$, we assume

$$
\mu_t = u \cdot w_t \quad \text{and} \quad \Sigma^{-1} = \frac{1}{\rho_1^2} \begin{pmatrix} \cos \alpha & \sin \alpha \\ -c \cdot \sin \alpha & c \cdot \cos \alpha \end{pmatrix}^T \begin{pmatrix} \cos \alpha & \sin \alpha \\ -c \cdot \sin \alpha & c \cdot \cos \alpha \end{pmatrix},
$$

(2.11)

where $u \in \mathbb{R}$, $c > 0$, and $\alpha \in [0, \pi/2]$. We use a wind vector which is averaged over the entire area, but the wind could also change locally. The motivation for writing $\Sigma$ in the given form comes from considering a coordinate transformation

$$
\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos \alpha & \sin \alpha \\ -c \cdot \sin \alpha & c \cdot \cos \alpha \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix},
$$

(2.12)

where the parameter $\alpha$ is the angle of rotation, and $c$ determines the degree of anisotropy, $c = 1$ corresponding to the isotropic case. $\rho_1$ is
a range parameter that determines the degree of interaction between spatial and temporal correlation. See Section 2.4.2 for an illustration of a kernel with the above parametrization.

The resulting model is non-stationary and incorporates anisotropy. Finally, we note that there are various other possible choices of parametrizations. For instance, a relatively simple model can be obtained by assuming

\[ \mu_t = 0 \quad \text{and} \quad \Sigma^{-1} = \frac{1}{\rho_1^2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \]  

(2.13)

that is no drift and an isotropic diffusion term. There is still spatio-temporal interaction, though, which implies that the model is not separable in the sense that (2.16) does not hold. We can simplify further and take not only \( \mu_t = 0 \), but also \( \Sigma = 0 \), leading to \( G_t \) being the identity matrix

\[ \xi_t = \phi \xi_{t-1} + \epsilon_t. \]  

(2.14)

This means that each point at time \( t - 1 \) only has an influence on itself at time \( t \), i.e., there is no spatio-temporal interaction and the model is separable.

### 2.2.3 Discussion of the model

**Propagator matrix** \( G_t \). Using a parametrized propagator matrix \( G_t \) in (2.6) has the obvious advantage that less parameters are needed than in the general case, in which each entry in the matrix has to be estimated, resulting in \( N^2 \) parameters. Moreover, in contrast to the general case, the parametric approach allows for making predictions at sites where no measurements are available, which is often of interest in applications.

**Space resolution consistency.** At first sight, it might be tempting to use a simpler parametrization of \( G_t \) not based on a convolution but of the form

\[ (G_t)_{ij} = \exp \left( -\frac{(d_{ij}/\rho_1)^2}{2} \right). \]  

(2.15)

However, such a model has the following important drawback. Assume, for instance, that a station \( i \) is surrounded by two neighbouring sites \( j \) and \( k \). Say that both stations \( j \) and \( k \) lie at the same distance from \( i \)
but in different directions. Consequently, $j$ and $k$ at time $t - 1$ exercise the same influence on $i$ at time $t$. If one adds an additional station $l$ very close to $k$, the joint influence of $k$ and $l$ at time $t - 1$ on site $i$ at time $t$ would then approximately be twice as big as the one of site $j$. This means that the distribution of the process at point $i$ depends on the number and the location of stations in the neighbourhood at which it has been observed. The convolution model, on the other hand, does not exhibit this drawback. Furthermore, the convolution model has the advantage that it is “space resolution consistent”, i.e., it retains approximately its temporal Markovian structure if one, or several, sites are removed from the domain. This does not hold true for the simpler vector autoregressive model as specified in (2.15).

**Space-time covariance structure.** In the following, let us turn to the spatio-temporal dependence structure of the latent process $\xi_t$. A random field $\xi_t(s), (s, t) \in \mathbb{R}^2 \times \mathbb{R}$ is said to have a separable covariance structure [Gneiting et al. (2007b)] if there exist purely spatial and purely temporal covariance functions $C_S$ and $C_T$, respectively, such that

$$\text{cov}(\xi_{t_1}(s_1), \xi_{t_2}(s_2)) = C_S(s_1, s_2) \cdot C_T(t_1, t_2).$$

(2.16)

The convolution based approach allows for nonseparable covariance structures, whereas the separable autoregressive model in (2.14) has a separable covariance structure.

**Extremal events.** For the data model as specified in equation (2.1), Hernández et al. (2009) showed that the distribution of the maxima is a Gumbel. If the focus lies on extremal events, other distributions, which have Fréchet maxima, can be used, for instance a $t-$distribution. The $t-$distribution is particularly attractive since it is a scale mixture of normal distributions. To be more specific, if $S_t$ has a $\chi^2_{df}$ distribution, then $W_t = x_t^T \beta + (\xi_t + \nu_t)/\sqrt{S_t/df}$ has a multivariate $t-$distribution. This means that the fitting algorithm introduced below can be extended to the $t-$distribution case by introducing an additional latent variable $S_t$. 
2.3 Fitting and prediction

Fitting is done using a Markov chain Monte Carlo method (MCMC), the Metropolis-Hastings algorithm [Metropolis et al. (1953), Hastings (1970)]. Concerning most parameters, it will be shown that the full conditionals are known distributions. Therefore, Gibbs sampling [Gelfand and Smith (1990)] can be used in these cases.

For convenience and later use, we combine the parameters characterizing the model into a vector \( \theta = (\lambda, \beta', \tau^2, \sigma^2, \rho_0, \vartheta')' \) and call them primary parameters. Our goal is to simulate from the joint posterior distribution of these parameters and the latent variables \( \xi = (\xi_1, \ldots, \xi_T), \xi_0, \) and \( W = (W_1, \ldots, W_T) \). We note that those \( W_t(s_i) \) that correspond to observed values above zero are known. In that case the full conditional distribution consists of a Dirac distribution at \( Y_t(s_i)^{1/\lambda} \). For handling the censored values and for allowing for missing values, we adopt a data augmentation approach [Smith and Roberts (1993)] as specified below in equation (2.19). See Section 2.3.1 for more details.

Assuming prior independence among the primary parameters and \( \xi_0 \), the prior distributions are specified as

\[
P[\lambda, \beta, \tau^2, \sigma^2, \rho_0, \rho_1, \alpha, c, \xi_0] \propto \frac{1}{\tau^2} \frac{1}{\sigma^2} P[\rho_0]P[\rho_1]P[u]P[c]P[\alpha] P[\xi_0|\sigma^2, \rho_0]
\]

(2.17)

with \( \xi_0 \) having a normal prior \( P[\xi_0|\sigma^2, \rho_0] = N(0, \sigma^2 V_{\rho_0}) \). Further, \( \rho_0 \) and \( \rho_1 \) have gamma priors with mean \( \mu_\rho \) and variance \( \sigma^2_\rho \). For \( c, \) we assume a gamma prior with mean 1 and variance 1, \( \alpha \) has a uniform prior on [0, \( \pi/2 \)], and \( u \) has a normal prior with mean 0 and variance 10^4. Further, we assume locally uniform priors on \( \log(\tau^2) \) and \( \log(\sigma^2) \) as well as for \( \phi, \lambda, \) and \( \beta \).

In our application, we choose to use informative priors for \( \rho_0 \) and \( \rho_1 \). It is known that in model-based geostatistics difficulties can arise when estimating the variance and scale parameters of the exponential covariogram [see, e.g., Warnes and Ripley (1987), Mardia and Watkins (1989), Diggle et al. (1998)]. For the geostatistical covariance model,
Chapter 2. A dynamic non-stationary spatio-temporal model
for short term prediction of precipitation

Zhang (2004) shows that the product of the two parameters can be estimated consistently, and Stein (1990) shows that it is the product of the two parameters that matters more than the individual parameters for spatial interpolation. Further, Berger et al. (2001) show that, at least in the simplest setting, the posterior of the range parameters is improper for most non-informative priors. Given these considerations, we think that using informative priors for the two range parameters $\rho_0$ and $\rho_1$ is appropriate. In our example, we chose priors with mean $\mu_\rho = 100$ and variance $\sigma_\rho^2 = 10$. We have tried different informative priors. The less informative they are, the worse are the mixing properties of the MCMC algorithm. In line with the results of Stein (1990) and Zhang (2004), we have made the experience that different choices of priors on these range parameters do not have a strong impact on the predictive performance of the model.

The posterior distribution is then proportional to

$$
\left( \frac{1}{\sigma^2} \right)^{\frac{N(T+1)}{2}+1} \left( \frac{1}{\tau^2} \right)^{\frac{NT}{2}+1} |V_{\rho_0}|^{-\frac{T+1}{2}} \prod_{Y_t(s_i) > 0} Y_t(s_i)^{1/\lambda - 1} \\
\cdot \exp \left( - \frac{1}{2} \sum_{t=1}^{T} \frac{1}{\tau^2} ||W_t - x_t^T \beta - \xi_t||^2 \right) \\
\cdot \exp \left( \frac{1}{\sigma^2} \left( \xi_t - \phi G_t \xi_{t-1} \right)^T V^{-1}_{\rho_0} \left( \xi_t - \phi G_t \xi_{t-1} \right) \right) \\
\cdot \exp \left( - \frac{1}{2} \frac{1}{\sigma^2} \xi_0^T V^{-1}_{\rho_0} \xi_0 \right) \cdot P[\rho_0] \cdot P[\theta] \cdot 1_{\{W_t(s_i) \leq 0 \ \forall \ i, t: \ Y_t(s_i) = 0\}}.
$$

(2.18)

The product in the first line is the Jacobian for the power transformation in (3.32). Note that missing observations do not cause any problem. If $Y_t(s_i)$ is missing, there is no respective term in the product nor a corresponding condition for the indicator function.

### 2.3.1 Full conditional distributions

In the following, we derive full conditional distributions for the individual parameters.
It is readily seen that the full conditional of $\beta$ is a multivariate normal distribution, and the full conditional distribution of $\phi$ is a normal distribution as well. The full conditionals of both $\sigma^2$ and $\tau^2$ are inverse gamma distributions.

For obtaining the full conditionals of $W_t$, we partition its components according to whether $Y_t(s_i)$ is above zero, equal to zero, or missing. Denote by $i_t^{[+]}$ those indices for which $Y_t(s_i) > 0$, by $i_t^{[0]}$ those with $Y_t(s_i) = 0$, and by $i_t^{[m]}$ the missing ones. The vector $W_t$ can then be partitioned into $W_t^{[+]}$, $W_t^{[0]}$, and $W_t^{[m]}$ accordingly. We remark that $W_t^{[0]}$ and $W_t^{[m]}$ are latent variables, whereas $W_t^{[+]}$ correspond to transformed observed values. In addition, $W_t^{[0]}$ has the restriction that all its values must be smaller than zero, $W_t^{[0]} \leq 0$. To facilitate understanding, we note that $W_t(s_i)$ can be written as

$$W_t(s_i) = W_t^{[+]}(s_i) = Y_t(s_i)^{1/\lambda}, \quad \text{if } Y_t(s_i) > 0,$$
$$= W_t^{[0]}(s_i), \quad \text{if } Y_t(s_i) = 0,$$
$$= W_t^{[m]}(s_i), \quad \text{if } Y_t(s_i) \text{ is missing.}$$

The full conditional of $W_t^{[m]}$ is then a multivariate normal distribution with mean and covariance

$$\mu_{W_t^{[m]}} = (x_t^T \beta + \xi_t)^{[m]} \quad \text{and} \quad \Sigma_{W_t^{[m]}} = \tau^2 \cdot I.$$ (2.20)

Similarly, the full conditional distribution of $W_t^{[0]}$ is a truncated multivariate normal distribution with mean and covariance

$$\mu_{W_t^{[0]}} = (x_t^T \beta + \xi_t)^{[0]} \quad \text{and} \quad \Sigma_{W_t^{[0]}} = \tau^2 \cdot I.$$ (2.21)

As mentioned before, the full conditional of $W_t^{[+]}$ is a Dirac distribution with point mass at $\left(Y_t^{[+]}\right)^{1/\lambda}$.

Concerning the latent variables $(\xi_0, \xi_1, \ldots, \xi_T)$, we note that conditional on $\theta$, $(\xi_t, W_t)$ is a linear Gaussian state space model. Therefore, a sample from the joint full conditional of $(\xi_0, \xi_1, \ldots, \xi_T)$ can be obtained using the forward filtering backward sampling (FFBS) algorithm.
proposed by Carter and Kohn (1994) and Frühwirth-Schnatter (1994). The forward filtering step corresponds to the Kalman filter [see, e.g., West and Harrison (1997) and Künsch (2001)].

Alternatively, one can also use single \( t \) updates. The full conditional of one \( \xi_t, 0 \leq t \leq T \), is a normal distribution \( N(\mu_{\xi_t}, \Sigma_{\xi_t}) \). In the case of the separable model, the mean \( \mu_{\xi_t} \) depends on \( \xi_{t-1} \) and \( \xi_{t+1} \), whereas the covariance matrix \( \Sigma_{\xi_t} \) does not depend on \( t \). This is convenient for simulation since its Cholesky decomposition has to be calculated only once in each update cycle. In contrast, in the sampling step of the FFBS algorithm, one has to calculate a Cholesky decomposition for each \( t \). The advantage that the FFBS algorithm mixes better than the single \( t \) update algorithm per update cycle, is outweighed by the fact that an update cycle of the single \( t \) update algorithm is a lot faster than one of the FFBS algorithm. Thus, more effective samples can be obtained with the single \( t \) update algorithm per time. In the case of the non-stationary anisotropic drift model, however, \( \Sigma_{\xi_t} \) in the single \( t \) update algorithm is not constant over time. Thus, a Cholesky decomposition needs to be computed for each \( t \) anyway, meaning that the FFBS algorithm is preferable.

In summary, we made the experience that it is recommendable to use single \( t \) updates for temporally stationary models where the covariance \( \Sigma_{\xi_t} \) of the full conditional of one \( \xi_t \) is constant over time. If \( \Sigma_{\xi_t} \) changes over time, we recommend using the FFBS algorithm.

For the remaining parameters, i.e., \( \rho_0, \vartheta \) (excluding \( \phi \)), and \( \lambda \), there is no apparent distribution family from which one can simulate. Metropolis steps will be used therefore. We note that the full conditional distribution of \( \lambda \) is proportional to

\[
\prod_{Y_t(s_i) > 0} \left( \frac{Y_t(s_i)^{1/\lambda - 1}}{\lambda} \right) \exp \left( -\frac{1}{2} \sum_{Y_t(s_i) > 0} \frac{1}{\tau^2} ||Y_t(s_i)^{1/\lambda} - x^T \beta - \xi_t||^2 \right).
\]

The parameter \( \lambda \) is sampled on the log-scale. This means that we first transform it to the log scale. Then a proposal is obtained by sampling from a normal distribution with the mean equal to the last value of the parameter. Thereafter, this proposal is accepted with a probability that is given by the usual Metropolis-Hasting algorithm [see, e.g., Chib...
and Greenberg (1995)].

Finally, $\rho_0$ and $\vartheta$ (excluding $\phi$) are sampled together. The full conditional is proportional to

$$
\exp \left( -\frac{1}{2\sigma^2} \left( \sum_{t=1}^T (\xi_t - \phi G_t \xi_{t-1})' V_{\rho_0}^{-1} (\xi_t - \phi G_t \xi_{t-1}) + \xi_t' V_{\rho_0}^{-1} \xi_t \right) \right) \cdot |V_{\rho_0}|^{-\frac{T+1}{2}}.
$$

(2.23)

### 2.3.2 Prediction

We consider predictions at new locations and / or times as well as predictions of areal averages. It turns out that in the case of areal averages, the Voronoi tessellation is again useful.

One way to obtain predictions is to augment the data $Y_{obs}$ with missing values at the locations or times where predictions are made. When doing so, the MCMC algorithm implicitly draws from the corresponding predictive distribution. See the previous Section 2.3.1 on how to handle missing values.

If one does not specify the points in space and time where predictions are to be made prior to model fitting, the predictive distribution of a new set of observations $Y^* = (Y_{t_1}^*(s_1^*), \ldots, Y_{t_k}^*(s_k^*))'$ is calculated as

$$
P[Y^*|Y_{obs}] = \int P[Y^*|\xi^*, \theta] P[\xi^*|\xi, \theta] P[\xi, \theta|Y_{obs}] d\xi^* d\xi d\theta
\approx \frac{1}{m} \sum_{i=1}^m \int P[Y^*|\xi^*, \theta^{(i)}] P[\xi^*|\xi^{(i)}, \theta^{(i)}] d\xi^* 
\approx \frac{1}{m} \sum_{i=1}^m P[Y^*|\xi^{*(i)}, \theta^{(i)}],
$$

(2.24)

where $Y_{obs}$ denotes the observed data, $\xi$ and $\xi^*$ the latent Gaussian process at the observed and predicted sites, respectively, and $\theta$ all the remaining parameters. Samples $\theta^{(i)}$ and $\xi^{(i)}$, $i = 1, \ldots, m$, from their posterior distribution are obtained by the MCMC algorithm, and $\xi^{*(i)}$ is sampled from $P[\xi^*|\xi^{(i)}, \theta^{(i)}]$. 
When $\xi^*$ is modeled at the same sites as $\xi$ but at different time points, the distribution $P[\xi^*|\xi^{(i)}, \theta^{(i)}]$ is Gaussian and readily obtained using (2.6).

In the case when predictions are made at unobserved sites $s \in S$ and time $t$, $P[\xi_t^*|\xi, \theta]$ can be calculated as described in the following. First, because of the temporal Markov property, $P[\xi_t^*|\xi_t, \theta]$ is equal to $P[\xi_t^*|\xi_{t-1}, \xi_t, \xi_{t+1}, \theta]$. This density is then obtained by considering the augmented model

$$
\begin{pmatrix}
\xi_t^* \\
\xi_t
\end{pmatrix} = \phi
\begin{pmatrix}
G_t \\
G_t^*
\end{pmatrix}
\xi_{t-1} + 
\begin{pmatrix}
\epsilon_t \\
\epsilon_t^*
\end{pmatrix},
\xi_{t+1} = \phi(H_{t+1}H_t^*)
\begin{pmatrix}
\xi_t^* \\
\xi_t
\end{pmatrix} + \epsilon_{t+1}.
$$

where $G_t^*$ is defined analogously to (2.7), $H_{t+1}$ and $H_t^*$ are obtained from the same approximations as in (2.5), and the covariances of $\epsilon_t$ and $\epsilon_t^*$ are as in (2.9). By (2.25), the conditional distribution of $\xi_t, \xi_t^*, \xi_{t+1}$ given $\xi_t-1$ is normal. Therefore also the conditional distribution of $\xi_t^*$ given $\xi_{t-1}, \xi_t, \xi_{t+1}$ is Gaussian. Its mean and covariance can be computed by noting that

$$
P[\xi_{t+1}^*|\xi_{t-1}, \xi_t, \xi_{t+1}, \theta] \propto P[\xi_{t+1}^*|\xi_{t-1}, \xi_t, \theta]P[\xi_{t+1}^*|\xi_{t-1}, \xi_t, \theta]
\propto P[\xi_{t+1}^*|\xi_{t-1}, \xi_t, \theta]P[\xi_t, \xi_t^*|\xi_{t-1}, \theta],
$$

and then completing the square in the exponent of the above expression.

In many cases, for instance when the focus lies on flooding, areal averages

$$
\bar{Y}_t(A^*) = \frac{1}{|A^*|} \int_{A^*} Y_t(s) ds
$$

of precipitation are of interest. If $Y_t(s)$ is observed on an irregular grid, one could first define a regular grid, then interpolate the non-observed grid points, and approximate the integral in (2.27) by a Riemann sum. However, since the regular grid usually becomes very large, this is computationally expensive. Instead, we propose to use the Voronoi tessellation once again to approximate the integral

$$
\bar{Y}_t(A^*) = \frac{1}{|A^*|} \int_{A^*} Y_t(s) ds \approx \frac{1}{|A^*|} \sum_{j=1}^{N} Y_t(s_j)|A_j \cap A^*|.
$$

(2.28)
Thereby, an adequate weight $|A_j \cap A^*|$ is given to each station. Samples from the predictive distribution of $\bar{Y}^{(A)}_t$ can be obtained by simulating $Y^{(i)}_t(s_j)$ from their predictive distribution and inserting them in (2.28).

We note that the areal prediction becomes deterministic if all $Y_t(s_j)$ consist of observed values. This means that uncertainty about values of $Y_t(s)$ at locations where no observations are made is implicitly ignored with the above approximation. This can be amended for by first making predictions at a few sites where no observations were made. Inserting additional unobserved sites can also be useful in other cases. For instance, if $A^*$ cuts off a substantial part of any $A_j$, i.e., $A_j \cap A^*$ is much smaller than $A_j$ but not empty, the areal prediction might be improved by replacing $Y_t(s_j)$ by the prediction of $Y_t$ at the center of gravity of $A_j \cap A^*$, or if the area $A^*$ is small and contains only a few stations, improved predictions of the areal average can be obtained by making predictions at a few additional points inside the area.

2.4 Application to short term prediction of precipitation

We apply the model to obtain short term forecasts of precipitation. Such forecasts are important, for instance, for agriculture and flooding. The traditional way for obtaining precipitation forecasts is the use of numerical weather prediction (NWP) models. NWP models solve complex, nonlinear equations emulating the dynamics of the atmosphere. Typically, NWP models require a lot of computational resources to run. Fitting our statistical model using the MCMC algorithm presented above is also computationally intensive. However, once the statistical model is fitted and assuming that the posterior of the primary parameters does not change (see Section 2.4.3 for more details), predictions are computationally a lot cheaper. Furthermore, the statistical model can be used in situations where there are no NWP models available or to obtain predictions at different temporal resolutions than the one at which the NWP model operates.
Chapter 2. A dynamic non-stationary spatio-temporal model for short term prediction of precipitation

2.4.1 The data

The data consists of three-hourly precipitation amounts collected by 26 stations around the Swiss Plateau from the beginning of December 2008 to the end of March 2009, making a total of 968 time periods. The data were provided by MeteoSwiss. We use the first three months, consisting of 720 time periods, for fitting the model. The remaining month March, consisting of 248 time periods, is set aside for model evaluation. The locations of these stations are shown in Figure 3.5. In Figure 3.6, a time series plot of the observed precipitation at one station (corresponding to the station with the acronym WYN in Figure...
3.5) and of the weighted areal average is shown. Concerning the latter, we take the weighted average over the entire spatial domain. Figure 2.3 shows the spatial distribution of the precipitation accumulated over time.

Figure 2.3: Illustration of the spatial distribution of precipitation. The circles display the cumulative rainfall amounts over time at the stations. The larger the circle and the darker the color, the higher is the cumulative precipitation amount. Both axes are in km.

The covariates consist of the x- and y-coordinates (km), altitude (m), temperature (°C), dew point (°C), and specific humidity (%). Specific humidity is the ratio of water vapor to dry air in a particular mass. It is expected to be positively related to precipitation. The dew point is the temperature to which a given parcel of humid air must be cooled, at constant barometric pressure, for water vapor to condense into water. Thus, the lower the dew point, the lower is the chance
for precipitation. However, specific humidity and dew point are considerably negatively correlated. This makes it unclear, a priori, what their joint relation to precipitation is like. Temperature, dew point, and specific humidity are predicted variables obtained from an NWP model called COSMO-2. From the same model, we also obtain wind predictions (speed is in m/s). Furthermore, predictions of the statistical model are evaluated by comparing them to precipitation forecasts from the same NWP. Having a high resolution with a grid spacing of 2.2 km, the NWP model is able to resolve convective dynamics. The NWP model produces predictions once a day for 24 hours ahead starting at 0:00UTC. After assimilation and computation, forecasts are available at around 1:30UTC. For all meteorological variables, we use values at approximately 1,000 m above ground. This is the height where we consider these variables to be most influential for precipitation. All covariates are centered and standardized to unit variance. Centering covariates around their means is used in order to avoid correlations of the regression coefficients with the intercept and to reduce posterior correlations.

2.4.2 Fitting and results

In the following, the non-stationary anisotropic model incorporating the wind as an external drift term (see Section 2.2) is fitted. In addition, we also fit a separable model. We simulate from the posterior distributions of these models as outlined in Section 2.3.

After the burn-in period consisting of 5,000 draws, 195,000 samples from the Markov chain were used to characterize posterior distributions. Convergence was monitored by inspecting trace plots.

In Table 2.1, we show posterior modes as well as 95% credible intervals for the different parameters of the non-stationary anisotropic drift model. The coefficients of the geographic coordinates are not significant. Specific humidity has a large positive coefficient. As expected, higher humidity implies more rainfall. The dew point is also positively related to precipitation. Higher temperatures, on the other hand, seem to imply less precipitation.

For interpreting the fitted parameters governing the convolution kernel ($\rho_1$, $c$, $\alpha$, and $u$), we illustrate in Figure 2.4 the convolution
Table 2.1: Posterior modes and 95% credible intervals for the non-stationary, anisotropic model with an external drift.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mode</th>
<th>2.5%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.05</td>
<td>-1.21</td>
<td>-0.93</td>
</tr>
<tr>
<td>X</td>
<td>-0.0473</td>
<td>-0.133</td>
<td>0.0541</td>
</tr>
<tr>
<td>Y</td>
<td>-0.0108</td>
<td>-0.0846</td>
<td>0.0531</td>
</tr>
<tr>
<td>Z</td>
<td>0.00347</td>
<td>-0.0169</td>
<td>0.0247</td>
</tr>
<tr>
<td>Temp</td>
<td>-0.717</td>
<td>-0.856</td>
<td>-0.583</td>
</tr>
<tr>
<td>Dew Point</td>
<td>0.406</td>
<td>0.187</td>
<td>0.601</td>
</tr>
<tr>
<td>Spec Hum</td>
<td>1.14</td>
<td>0.949</td>
<td>1.33</td>
</tr>
<tr>
<td>λ</td>
<td>1.58</td>
<td>1.54</td>
<td>1.62</td>
</tr>
<tr>
<td>τ^2</td>
<td>0.0685</td>
<td>0.0451</td>
<td>0.0943</td>
</tr>
<tr>
<td>σ^2</td>
<td>1.04</td>
<td>0.953</td>
<td>1.17</td>
</tr>
<tr>
<td>ρ₀</td>
<td>92</td>
<td>86.4</td>
<td>97.9</td>
</tr>
<tr>
<td>φ</td>
<td>0.000159</td>
<td>0.000147</td>
<td>0.00017</td>
</tr>
<tr>
<td>ρ₁</td>
<td>93.6</td>
<td>88.1</td>
<td>99.4</td>
</tr>
<tr>
<td>c</td>
<td>4.1</td>
<td>3.61</td>
<td>4.63</td>
</tr>
<tr>
<td>α</td>
<td>0.704</td>
<td>0.658</td>
<td>0.777</td>
</tr>
<tr>
<td>u</td>
<td>0.879</td>
<td>0.645</td>
<td>1.1</td>
</tr>
</tbody>
</table>

The kernel over the region where the stations lie. The parameters ρ₁, c, α, and u are taken at their posterior mode. The plot is interpreted as follows. The height of the kernel is the level of influence that ξ_{t-1}(s') at location s' has on ξ_{t}(s) at location s as a function of s' – s. In other words, the colors represent the lag-1 influence of the other stations on the station Wynau which is used as origin in the plot. The white arrow represents the drift vector \( \mathbf{\mu}_t = u \cdot \mathbf{w}_t \) at time \( t = 429 \), \( \mathbf{w}_t \) being the wind vector. Note that that this transport vector changes over time thus causing temporal non-stationarity. The time \( t = 429 \) illustrates a meteorological situation with the typically predominant southwestern wind direction.

With c and α being approximately 4 and 0.7, we observe anisotropy along the south-east north-west direction. This corresponds to the topography of the region, as the area containing a majority of the
Figure 2.4: Illustration of the convolution kernel at time $t = 429$. The colors indicate the lag-1 influence of the other stations on the station Wynau. The white arrow represents the drift caused by a south-west wind at this time point. The dots represent the observation stations. The axes are in km.

stations lies between two mountain ranges: the Jura to the north-west and the Alps to the south-east. Correlations are expected to be higher along the flat part between these two mountain ranges.

Furthermore, the plot shows how the external drift shifts the convolution kernel. Apparently, the southwestern neighbor (Bern) has the highest influence on Wynau in this situation with wind coming from the southwest. Gneiting et al. (2006) observe a similar phenomenon in wind speed data over the U.S. Pacific Northwest where there is also a predominant wind direction causing asymmetric cross-correlations.

2.4.3 Short term prediction of precipitation

In the following, we apply the fitted models to produce short term predictions of precipitation. As mentioned before, we have fitted the model to the first 720 time periods from December 2008 to February 2009. From this we obtain posterior distributions for the primary pa-
2.4. Application to short term prediction of precipitation

Parameters. Predictions for the time periods in March that were set aside are obtained as described in the following.

Ideally, one would run the full MCMC algorithm at each time point, including all data up to the point, and obtain predictive distributions from this. However, since this is rather time consuming, we make the following approximation. We assume that the posterior distribution of the primary parameters given $Y_{1:t} = \{Y_1, \ldots, Y_t\}$ is the same for all $t \geq 720$. That is we neglect the additional information that the observations in March give about the primary parameters. In practice, this means that posterior distributions of the primary parameters are calculated only once, namely on the dataset from December 2008 to February 2009.

For each time $t \geq 720$, we make up to 8 steps ahead forecasts. I.e., we sample from the predictive distribution of $Y_{t+k}^*$, $k = 1, \ldots, 8$, given $Y_{1:t} = \{Y_1, \ldots, Y_t\}$ and given the posterior of the primary parameters based on the data from December 2008 to February 2009. Since the NWP produces forecasts for the three meteorological covariates once a day, for each prediction time $t + k$, the forecasts made at 0:00UTC of the same day are used. Sampling from the predictive distribution consists of imputing the augmented data $W$ and sampling from the latent process $\xi$. These two steps are done as described in Section 2.3. To generate one sample from the predictive distribution takes around 3.5 seconds on an AMD Athlon(tm) 64 X2 Dual Core Processor 5600+ with a 2900 MHz CPU clock rate. We use 200 samples to characterize each predictive distribution.

The assumption that the posterior of the primary parameters does not change may be questionable over longer time periods and when one moves away from the time period from which data is used to obtain the posterior distribution. But since all our data lies in the winter season, we think that this assumption is reasonable. If longer time periods are considered, one could use sliding training windows or model the primary parameters as evolving dynamically over time. One can also investigate how the predictive performance deteriorates with increasing lags between predictions and last time point from which data is used to fit the model.

In addition to the separable model and the non-stationary anisotropic
Chapter 2. A dynamic non-stationary spatio-temporal model for short term prediction of precipitation

Figure 2.5: Comparison of statistical models. The continuous ranked probability score (CRPS) of forecasts versus number of consecutive time periods for which predictions are made is shown. On the left are CRPSs of station specific forecasts and on the right are CRPSs of areal forecasts. “NoAR” denotes the model without an autoregressive term, “SAR” the one with a separable covariance structure, and “ConvAR” the convolution based non-stationary anisotropic drift model. All three models include the covariates described in Section 2.4.1. A convolution based model without including covariates (“ConvAR No Cov”) is also fitted. The unit of the CRPS is mm.

In order to assess the performance of the probabilistic predictions, we fit a model with no autoregressive term, i.e., with $\phi = 0$. Further, to assess how much information stems from the three meteorological covariates (temperature, dew point, and specific humidity) and how much from the dynamic spatio-temporal model, we also fit the non-stationary anisotropic drift model without including these covariates. For each model, we calculate pointwise predictions for the individual stations and also predictions for the areal average. The latter are obtained using the Voronoi tessellation as described in Section 2.3.2.
we use the continuous ranked probability score (CRPS) [Matheson and Winkler (1976)]. The CRPS is a strictly proper scoring rule [Gneiting and Raftery (2007)] that assigns a numerical value to probabilistic forecasts and assesses calibration and sharpness simultaneously [Gneiting et al. (2007a)]. It is defined as

$$CRPS(F, y) = \int_{-\infty}^{\infty} (F(x) - 1_{\{y \leq x\}})^2 dx,$$

where $F$ is the predictive cumulative distribution function, $y$ is the observed realization, and $1$ is an indicator function. It can be equivalently calculated as

$$CRPS(F, y) = E_F|Y - y| - \frac{1}{2} E_F|Y - Y'|,$$

where $Y$ and $Y'$ are independent random variables with distribution $F$. If a sample $Y^{(1)}, \ldots, Y^{(m)}$ from $F$ is available, it can be approximated by

$$\frac{1}{m} \sum_{i=1}^{m} |Y^{(i)} - y| - \frac{1}{2m^2} \sum_{i,j=1}^{m} |Y^{(i)} - Y^{(j)}|.$$

In Figure 3.7, the average CRPS of the pointwise predictions and the areal predictions are plotted versus lead times. In the left plot, the mean is taken over all stations and time periods, whereas the areal version is an average over all time periods. Predictions $Y^*_{t+k}, k = 1, \ldots, 8$ for the next 8 time steps are made at each time point $t$. We recall that the NWP model produces predictions for 8 consecutive periods once a day at midnight. For simplicity, potential diurnal variation in the accuracy of the predicted covariates is ignored.

We see that the non-stationary anisotropic drift model (“ConvAR”) has clearly the best performance among the three models. In particular, the non-separable convolution based model performs better than the simpler separable spatio-temporal model (“SAR”). Not surprisingly, the model without temporal dependency (“NoAR”) performs worse than the other two models. Comparing the “ConvAR” model, the non-stationary convolution model without covariates (“ConvAR No Cov”),
and the “NoAR” model, we see that the main source of predictive performance at small lead times are not the covariates but the dynamic spatio-temporal model. In the areal case, the non-stationary convolution model without covariates even outperforms the simple autoregressive model including covariates at small lead times. With increasing lead time, the meteorological covariates contribute more to the predictive performance and the dynamic spatio-temporal model becomes less important.

**Figure 2.6:** Comparison of statistical and NWP model. The mean absolute error (MAE) of forecasts versus lead time is shown. Lead time also corresponds to the time of day. The left panel shows MAEs of station specific forecasts averaged over time and the stations, and the right panel shows MAEs of areal forecasts averaged over time. “ConvAR” denotes the convolution based non-stationary anisotropic drift model and “NWP” the NWP model. The bold lines show the results when excluding March 24, 2009. The unit of the MAE is mm.

We also compare the performance of the predictions from the non-stationary anisotropic drift model with predictions obtained from the NWP model. Since the NWP model produces deterministic forecasts, we use the mean absolute error (MAE). In order to make the compari-
2.4. Application to short term prediction of precipitation

son fair, we first reduce the statistical distributional forecast to a point forecast by taking the median (see Gneiting (2011) on why this is a reasonable choice). As mentioned, the NWP model produces predictions once a day starting at 0:00UTC. Predictions are then made for eight consecutive time periods corresponding to 24 h ahead. This means that the time of day also corresponds to the lead time. This is in contrast to the above comparison of the different statistical models where 8 step ahead predictions were made at all time periods.

Table 2.2: Comparison of statistical and NWP model. The mean absolute error (MAE) averaged over all days and lead times is reported. "ConvAR" denotes the convolution based non-stationary anisotropic drift model and "NWP" the NWP model. The unit of the MAE is mm.

<table>
<thead>
<tr>
<th></th>
<th>ConvAR</th>
<th>NWP</th>
<th>Areal ConvAR</th>
<th>Areal NWP</th>
</tr>
</thead>
<tbody>
<tr>
<td>March 2009</td>
<td>0.41</td>
<td>0.46</td>
<td>0.35</td>
<td>0.32</td>
</tr>
<tr>
<td>Excl. March 24</td>
<td>0.36</td>
<td>0.43</td>
<td>0.29</td>
<td>0.31</td>
</tr>
</tbody>
</table>

In Figure 2.6, the mean absolute error (MAE) of forecasts versus lead time, or time of day, respectively, is shown. In addition, in Table 2.2, we report MAEs averaged over all lead times. Note that there is one particular day (March 24) when heavy rainfall occurred shortly after 0:00UTC. We report results including (thin lines) and excluding (bold lines) this day.

Table 2.2 shows that overall the statistical model outperforms the NWP on a stationwise base. When considering the areal average, the two models perform similarly. Depending on whether March 24 is included or not, the NWP or the statistical model has a slightly lower average MAE.

Furthermore, Figure 2.6 shows that March 24 considerably affects the performance of the one- and two-step ahead predictions of the statistical model as well as the stationwise performance of the NWP model. When excluding this day, the corresponding MAEs are considerably lower. This shows a typical behaviour of our model and statistical
models in general: they perform well when, at the time of prediction, the major phenomena (advective fronts) are already observable. In this case, the spatio-temporal statistical model can extrapolate the space-time dynamics of the rainfall process into the future.

Earlier studies have shown that nowcasting methods, including statistical approaches, perform usually better at short lead times (up to one day) while NWP have higher predictive skills at medium ranges [see Kober et al. (2012) or Little et al. (2009)]. Our results are in line with these findings in the sense that all lead times used in our application are still in the range of what is considered “short” lead times. However, our model is not just based on past precipitation observations but also on other predicted meteorological variables.

\section*{2.5 Conclusions}

A hierarchical Bayesian spatio-temporal model is presented. Incorporating physical knowledge, the dynamic model is non-stationary, anisotropic, and allows for non-separable covariance structures. It incorporates a drift term that depends on a wind vector. At the data stage, the model determines the probability of rainfall and the rainfall amount distribution together. This is done using a normal variable that depends linearly on covariates. The model is fitted using Markov chain Monte Carlo (MCMC) methods and applied to obtain short term precipitation forecasts. It performs better than a separable, stationary, and isotropic model, and it performs comparably to a deterministic numerical weather prediction model and has the advantage that it quantifies prediction uncertainty.

Even though we have applied the model to prediction of precipitation, it can also be used to predict or interpolate other meteorological quantities of interest.

Future research could focus on adapting the model so that it can be applied to spatially highly resolved data. Using Markov random fields [Rue and Held (2005), Lindgren et al. (2011)] for the innovation process $\epsilon_t$ might be a potential direction. Alternatively, a dimension reduction approach could be examined, cf. Banerjee et al. (2008). For instance, Sigrist et al. (2012b) approximate an advection-diffusion SPDE to cope
with large data sets. Further, the model can be extended by addi-
tionally relaxing some assumptions. For instance, the parameters \( \sigma^2, \phi, \rho_0, \rho_1, \) and \( \lambda \) were assumed to be constant over time. Assuming pe-
periodicity, Fourier harmonics could be used to model parameters that
vary seasonally during the year. Alternatively, the parameters could
evolve dynamically over time according to an equation of the form
\( \vartheta_t = \vartheta_{t-1} + N(0, \sigma^2_{\vartheta}) \).
Increasingly larger data sets of processes in space and time ask for statistical models and methods that can cope with such data. We show that solutions of stochastic advection-diffusion partial differential equations (SPDEs) provide a flexible model class for spatio-temporal processes which is computationally feasible also for large data sets. The solution of the SPDE has in general a nonseparable covariance structure. Furthermore, its parameters can be physically interpreted as explicitly modeling phenomena such as transport and diffusion that occur in many natural processes in diverse fields ranging from environmental sciences to ecology. In order to obtain computationally efficient statistical algorithms we use spectral methods to solve the SPDE. This has the advantage that approximation errors do not accumulate over time, and that in the spectral space the computational cost grows linearly with the dimension, the total computational costs of Bayesian or frequentist inference being dominated by the fast Fourier transform. The proposed model is applied to postprocessing of precipitation forecasts from a numerical weather prediction model for northern Switzerland. In contrast to the raw forecasts from the numerical model, the post-
processed forecasts are calibrated and quantify prediction uncertainty. Moreover, they outperform the raw forecasts.

3.1 Introduction

Space-time data arise in many applications, see Cressie and Wikle (2011) for an introduction and an overview. Increasingly larger space-time data sets are obtained, for instance, from remote sensing satellites or deterministic physical models such as numerical weather prediction (NWP) models. Statistical models are needed that can cope with such data.

As Wikle and Hooten (2010) point out, there are two basic paradigms for constructing spatio-temporal models. The first approach is descriptive and follows the traditional geostatistical paradigm, using joint space-time covariance functions (Cressie and Huang, 1999; Gneiting, 2002; Ma, 2003; Wikle, 2003; Stein, 2005; Paciorek and Schervish, 2006). The second approach is dynamic and combines ideas from time-series and spatial statistics (Solna and Switzer, 1996; Wikle and Cressie, 1999; Huang and Hsu, 2004; Xu et al., 2005; Gelfand et al., 2005; Johannesson et al., 2007; Sigrist et al., 2012).

Even for purely spatial data, developing methodology which can handle large data sets is an active area of research. Banerjee et al. (2004) refer to this as the “big n problem”. Factorizing large covariance matrices is not possible without assuming a special structure or using approximate methods. Using low rank matrices is one approach (Nychka et al., 2002; Banerjee et al., 2008; Cressie and Johannesson, 2008; Stein, 2008; Wikle, 2010). Other proposals include using Gaussian Markov random-fields (GMRF) (Rue and Tjelmeland, 2002; Rue and Held, 2005; Lindgren et al., 2011) or applying tapering (Furrer et al., 2006) so that one obtains sparse precision or covariance matrices, respectively, for which calculations can be done efficiently. Another proposed solution is to approximate the likelihood so that it can be evaluated faster (Vecchia, 1988; Stein et al., 2004; Fuentes, 2007; Eidsvik et al., 2012). Royle and Wikle (2005) and Paciorek (2007) use Fourier functions to reduce computational costs.

In a space-time setting, the situation is the same, if not worse:
3.1. Introduction

one runs into a computational bottleneck with high dimensional data since the computational costs to factorize dense $NT \times NT$ covariance matrices are $O((NT)^3)$, $N$ and $T$ being the number of points in space and time, respectively. Moreover, specifying flexible and realistic space-time covariance functions is a nontrivial task.

In this paper, we follow the dynamic approach and study models which are defined through a stochastic advection-diffusion partial differential equation (SPDE). This has the advantage of providing physically motivated parametrizations of space-time covariances. We show that when solving the SPDE using Fourier functions, one can do computationally efficient statistical inference. In the spectral space, computational costs for the Kalman filter and backward sampling algorithms are of order $O(NT)$. As we show, roughly speaking, this computational efficiency is due to the temporal Markov property, the fact that Fourier functions are eigenfunctions of the spatial differential operators, and the use of some matrix identities. The overall computational costs are then determined by the ones of the fast Fourier transform (FFT) (Cooley and Tukey, 1965) which are $O(TN \log N)$. In addition, computational time can be further reduced by running the $T$ different FFTs in parallel.

Defining Gaussian processes through stochastic differential equations has a long history in statistics going back to early works such as Whittle (1954), Heine (1955), and Whittle (1962). Later works include Jones and Zhang (1997) and Brown et al. (2000). Recently, Lindgren et al. (2011) showed how a certain class of SPDEs can be solved using finite elements to obtain parametrizations of spatial GMRF.

Spectral methods for solving partial differential equation are well established in the numerical mathematics community (see, e.g., Gottlieb and Orszag (1977), Folland (1992), or Haberman (2004)). In contrast, statistical models have different requirements and goals, since the (hyper-)parameters of an (S)PDE are not known a priori and need to be estimated. Spectral methods have also been used in spatio-temporal statistics, mostly for approximating or solving deterministic integro-difference equations (IDEs) or PDEs. Wikle and Cressie (1999) introduce a dynamic spatio-temporal model obtained from an IDE that is approximated using a reduced-dimensional spectral basis. Extending

The novel features of our work are the following. While spectral methods have been used for approximating deterministic IDEs and PDEs in the statistical literature, there is no article, to our knowledge, that explicitly shows how one obtains a space-time Gaussian process by solving an advection-diffusion SPDE using the real Fourier transform. Moreover, we present computationally efficient algorithms, for doing statistical inference, which use the fast Fourier transform and the Kalman filter. The computational burden can be additionally alleviated by applying dimension reduction. We also give a bound on the accuracy of the approximate solution. In the application, we postprocess precipitation forecasts, explicitly modeling spatial and temporal variation. The idea is that the spatio-temporal model not only accounts for dependence, but also captures and extrapolates dynamically an error term of the NWP model in space and time.

The remainder of this paper is organized as follows. Section 3.2 introduces the continuous space-time Gaussian process defined through the advection-diffusion SPDE. In Section 3.3, it is shown how the solution of the SPDE can be approximated using the two-dimensional real Fourier transform, and we give convergence rates for the approximation. Next, in Section 3.4, we show how one can do computationally efficient inference. In Section 3.5, the spatio-temporal model is used as part of a hierarchical Bayesian model, which we then apply for postprocessing of precipitation forecasts.

All the methodology presented in this article is implemented in the R package spate (see Sigrist et al. (2012a)).
3.2 A Continuous Space-Time Model: The Advection-Diffusion SPDE

In one dimension, a fundamental process is the Ornstein-Uhlenbeck process which is governed by a relatively simple stochastic differential equation (SDE). The process has an exponential covariance function and its discretized version is the famous AR(1) model. In the two dimensional spatial case, Whittle (1954) argues convincingly that the process with a Whittle correlation function is an “elementary” process (see Section 3.2.2 for further discussion). If the time dimension is added, we think that the process defined through the stochastic partial differential equation (SPDE) in (3.1) has properties that make it a good candidate for an “elementary” spatio-temporal process. It is a linear equation that explicitly models phenomena such as transport and diffusion that occur in many natural processes ranging from environmental sciences to ecology. This means that, if desired, the parameters can be given a physical interpretation. Furthermore, if some parameters equal zero (no advection and no diffusion), its covariance structure reduces to a separable one with an AR(1) structure over time and a certain covariance structure over space.

The advection-diffusion SPDE, also called transport-diffusion SPDE, is given by

$$\frac{\partial}{\partial t} \xi(t, s) = -\mathbf{\mu}' \nabla \xi(t, s) + \nabla \cdot \mathbf{\Sigma} \nabla \xi(t, s) - \zeta \xi(t, s) + \epsilon(t, s), \quad (3.1)$$

with \( s = (x, y)' \in \mathbb{R}^2 \), where \( \nabla = \left( \frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right)' \) is the gradient operator, and, for a vector field \( \mathbf{F} = (F^x, F^y)' \), \( \nabla \cdot \mathbf{F} = \frac{\partial F^x}{\partial x} + \frac{\partial F^y}{\partial y} \) is the divergence operator. \( \epsilon(t, s) \) is a Gaussian process that is temporally white and spatially colored. See Section 3.2.2 for a discussion on the choice of the spatial covariance function. Heine (1955) and Whittle (1963) introduced and analyzed SPDEs of similar form as in (3.1). Jones and Zhang (1997) also investigated SPDE based models. Furthermore, Brown et al. (2000) obtained such an advection-diffusion SPDE as a limit of stochastic integro-difference equation models. Without giving any concrete details, Lindgren et al. (2011) suggested that this SPDE can be used in connection with their GMRF method.
The SPDE has the following interpretation. Heuristically, an SPDE specifies what happens locally at each point in space during a small time step. The first term $\mu' \nabla \xi(t,s)$ models transport effects (called advection in weather applications), $\mu = (\mu_x, \mu_y)' \in \mathbb{R}^2$ being a drift or velocity vector. The second term, $\nabla \cdot \Sigma \nabla \xi(t,s)$, is a diffusion term that can incorporate anisotropy. If $\Sigma$ is the identity matrix, this term reduces to the divergence ($\nabla \cdot$) of the gradient ($\nabla$) which is the ordinary Laplace operator $\nabla \cdot \nabla = \Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$. The third term $-\zeta \xi(t,s)$, $\zeta > 0$, diminishes $\xi(t,s)$ at a constant rate and thus accounts for damping. Finally, $\epsilon(t,s)$ is a source-sink or stochastic forcing term, also called innovation term, that can be interpreted as describing, amongst others, convective phenomena in precipitation modeling applications.

Concerning the diffusion matrix $\Sigma$, we suggest the following parametrization

$$\Sigma^{-1} = \frac{1}{\rho_1^2} \begin{pmatrix} \cos \psi & \sin \psi \\ -\gamma \cdot \sin \psi & \gamma \cdot \cos \psi \end{pmatrix}^T \begin{pmatrix} \cos \psi & \sin \psi \\ -\gamma \cdot \sin \psi & \gamma \cdot \cos \psi \end{pmatrix}, \quad (3.2)$$

where $\rho_1 > 0$, $\gamma > 0$, and $\psi \in [0, \pi/2]$. The parameters are interpreted as follows. $\rho_1$ acts as a range parameter and controls the amount of diffusion. The parameters $\gamma$ and $\psi$ control the amount and the direction of anisotropy. With $\gamma = 1$, isotropic diffusion is obtained.

Figure 3.1 illustrates the SPDE in (3.1) and the corresponding PDE without the stochastic innovation term. The top row shows a solution to the PDE which corresponds the deterministic part of the SPDE that is obtained when there is no stochastic term $\epsilon(t,s)$. The figure shows how the initial state in the top-left plot gets propagated forward in time. The drift vector points from north-east to south-west and the diffusive part exhibits anisotropy in the same direction. A $100 \times 100$ grid is used and the PDE is solved in the spectral domain using the method described below in Section 3.3. There is a fundamental difference between the deterministic PDE and the probabilistic SPDE. In the first case, a deterministic process is modeled directly. In the second case, the SPDE defines a stochastic process. Since the operator is linear and the input Gaussian, this process is a Gaussian process whose covariance function is implicitly defined by the SPDE. The bottom row of Figure 3.1 shows one sample from this Gaussian process. The same...
3.2. A Continuous Space-Time Model: The Advection-Diffusion SPDE

Figure 3.1: Illustration of the SPDE in (3.1) and the corresponding PDE. The top row illustrates a solution to the PDE which corresponds the deterministic part of the SPDE without stochastic term $\epsilon(t, s)$. The bottom row shows one sample from the distribution specified by the SPDE with a fixed initial condition. The drift vector points from north-east to south-west and the diffusive part exhibits anisotropy in the same direction. The same parameters are used for both the PDE and the SPDE: $\zeta = -\log(0.99)$, $\rho_1 = 0.06$, $\gamma = 3$, $\psi = \pi/4$, $\mu_x = -0.1$, $\mu_y = -0.1$, and for the stochastic innovations: $\rho_0 = 0.05$, $\sigma^2 = 0.7^2$. The color scales are different in different panels.
initial state as in the deterministic example is used, i.e., we use a fixed initial state. Except for the stochastic part, the same parameters are used for both the PDE and the SPDE. For the innovations \( \epsilon(t, s) \), we choose a Gaussian process that is temporally independent and spatially structured according to the Matérn covariance function with smoothness parameter 1. Again, the drift vector points from north-east to south-west and the diffusive part exhibits anisotropy in the same direction.

Note that the use of this spatio-temporal Gaussian process is not restricted to situations where it is a priori known that phenomena such as transport and diffusion occur. In the one dimensional case, it is common to use the AR(1) process in situations where it is not a priori clear whether the modeled process follows the dynamic of the Ornstein-Uhlenbeck SDE. In two dimensions, the same holds true for the process with the Whittle covariance function, and even more for the process having an exponential covariance structure. Having this in mind, even though the SPDE in (3.1) is physically motivated, it can be used as a general spatio-temporal model. As the case may be, the interpretation of the parameters can be more or less straightforward.

### 3.2.1 Spectral Density and Covariance Function

As can be shown using the Fourier transform (see, e.g., Whittle (1963)), if the innovation process \( \epsilon(t, s) \) is stationary with spectral density \( \tilde{f}(k) \), the spectrum of the stationary solution \( \xi(t, s) \) of the SPDE (3.1) is

\[
    f(\omega, k) = \tilde{f}(k) \frac{1}{2\pi} \left( (k' \Sigma k + \zeta)^2 + (\omega + \mu' k)^2 \right)^{-1},
\]

where \( k \) and \( \omega \) are spatial wavenumbers and temporal frequencies. The covariance function \( C(t, s) \) of \( \xi(t, s) \) is then given by

\[
    C(t, s) = \int f(\omega, k) \exp(i \cdot t\omega) \exp(i \cdot s' k) dk d\omega
    = \int \tilde{f}(k) \frac{\exp\left( -i \cdot \mu' k t - (k' \Sigma k + \zeta)|t| \right)}{2(k' \Sigma k + \zeta)} \exp(i \cdot s' k) dk,
\]

where \( i \) denotes the imaginary number \( i^2 = -1 \), and the integration over the temporal frequencies \( \omega \) follows from the calculation of the
characteristic function of the Cauchy distribution (Abramowitz and Stegun, 1964). The spatial integral above has no closed form solution but can be computed approximately by numerical integration.

Since, in general, the spectrum does not factorize into a temporal and a spatial component, we see that $\xi(t, s)$ has a non-separable covariance function (see Gneiting et al. (2007b) for a definition of separability). The model reduces to a separable one, though, when there is no advection and diffusion, i.e., when both $\mu$ and $\Sigma$ are zero. In this case, the covariance function is given by $C(t, s) = \frac{1}{2\zeta} \exp (-\zeta |t|) C(s)$, where $C(s)$ denotes the spatial covariance function of the innovation process.

### 3.2.2 Specification of the Innovation Process

As mentioned before, it is assumed that the innovation process is white in time and spatially colored. In principle, one can choose any spatial covariance function such that the covariance function in (3.4) is finite at zero. Note that if $\tilde{f}(k)$ is integrable, then $f(\omega, k)$ is also integrable. Similarly as Lindgren et al. (2011), we opt for the most commonly used covariance function in spatial statistics: the Matérn covariance function (see Handcock and Stein (1993), Stein (1999)). Since in many applications the smoothness parameter is not estimable, we further restrict ourselves to the Whittle covariance function. This covariance function is of the form $\sigma^2 d/\rho_0 K_1 (d/\rho_0)$ with $d$ being the Euclidean distance between two points and $K_1 (d/\rho_0)$ being the modified Bessel function of order 1. It is called after Whittle (1954) who introduced it and argued convincingly that it “may be regarded as the ’elementary’ correlation in two dimensions, similar to the exponential in one dimension.”. It can be shown that the stationary solution of the SPDE

$$ \left( \nabla \cdot \nabla - \frac{1}{\rho_0^2} \right) \epsilon(t, s) = W(t, s), \quad (3.5) $$

where $W(t, s)$ is a zero mean Gaussian white noise field with variance $\sigma^2$, has the Whittle covariance function in space. From this, it follows
that the spectrum of the process $\epsilon(t, s)$ is given by

$$
\tilde{f}(k) = \frac{\sigma^2}{(2\pi)^2} \left( k' k + \frac{1}{\rho_0^2} \right)^{-2}, \quad \rho_0 > 0, \quad \sigma > 0.
$$

(3.6)

The parameter $\sigma^2$ determines the marginal variance of $\epsilon(t, s)$, and $\rho_0$ is a spatial range parameter.

### 3.2.3 Relation to an Integro-Difference Equation

Assuming discrete time steps with lag $\Delta$, Brown et al. (2000) consider the following integro-difference equation (IDE)

$$
\xi(t, s) = \exp \left( -\Delta \zeta \right) \int_{\mathbb{R}^2} h(s - s') \xi(t - \Delta, s') ds' + \epsilon(t, s), \quad s \in \mathbb{R}^2,
$$

(3.7)

with a Gaussian redistribution kernel

$$
h(s - s') = \frac{1}{2\pi|2\Delta \Sigma|^{1/2}} \exp \left( -(s - s' - \Delta \mu)^T (2\Delta \Sigma)^{-1} (s - s' - \Delta \mu)/2 \right),
$$

$\epsilon(t, s)$ being temporally independent and spatially dependent. They show that in the limit $\Delta \to 0$, the solution of the IDE and the one of the SPDE in (3.1) coincide. The IDE is interpreted as follows: the convolution kernel $h(s - s')$ determines the weight or the amount of influence that a location $s'$ at previous time $t - \Delta$ has on the point $s$ at current time $t$. This IDE representation provides an alternative way of interpreting the SPDE model and its parameters. Storvik et al. (2002) show under which conditions a dynamic model determined by an IDE as in (3.7) can be represented using a parametric joint space-time covariance function, and vice versa. Based on the IDE in (3.7), Sigrist et al. (2012) construct a spatio-temporal model for irregularly spaced data and apply it to obtain short term predictions of precipitation.

### 3.3 Solution in the Spectral Space

Solutions $\xi(t, s)$ of the SPDE (3.1) are defined in continuous space and time. In practice, one needs to discretize both space and time. The resulting vector of $NT$ space-time points is in general of large dimension.
This makes statistical inference, be it frequentist or Bayesian, computationally difficult to impossible. However, as we show in the following, solving the SPDE in the spectral space alleviates the computational burden considerably and allows for dimension reduction, if desired.

Heuristically speaking, spectral methods (Gottlieb and Orszag, 1977; Cressie and Wikle, 2011, Chapter 7) approximate the solution $\xi(t, s)$ by a linear combination of deterministic spatial functions $\phi_j(s)$ with random coefficients $\alpha_j(t)$ that evolve dynamically over time:

$$
\xi^K(t, s) = \sum_{j=1}^{K} \alpha_j(t) \phi_j(s) = \phi(s)' \alpha(t),
$$

(3.8)

where $\phi(s) = (\phi_1(s), \ldots, \phi_K(s))'$ and $\alpha(t) = (\alpha_1(t), \ldots, \alpha_K(t))'$. To be more specific, we use Fourier functions

$$
\phi_j(s) = \exp \left( i \cdot k_j' s \right),
$$

(3.9)

where $k_j = (k_{jx}, k_{jy})'$ is a spatial wavenumber.

The advantages of using Fourier functions for solving linear, deterministic PDEs are well known, see, e.g., Pedlosky (1987). First, differentiation in the physical space corresponds to multiplication in the spectral space. In other words, Fourier functions are eigenfunctions of the spatial differential operator. Instead of approximating the differential operator in the physical space and then worrying about approximation errors, one just has to multiply in the spectral space, and there is no approximation error of the operator. In addition, one can use the FFT for efficiently transforming from the physical to the spectral space, and vice versa.

Proposition 1 shows that Fourier functions are also useful for the stochastic PDE (3.1): if the initial condition and the innovation process are in the space spanned by a finite number of Fourier functions, then the solution of the SPDE (3.1) remains in this space for all times and can be given in explicit form.

**Proposition 1.** Assume that the initial state and the innovation terms are of the form

$$
\xi^K(0, s) = \phi(s)' \alpha(0), \quad \epsilon^K(t, s) = \phi(s)' \tilde{\epsilon}(t)
$$

(3.10)
where $\phi(s) = (\phi_1(s), \ldots, \phi_K(s))'$, $\phi_j(s)$ is given in (3.9), $\alpha(0) \sim N\left(0, \text{diag}\left(\tilde{f}_0(k_j)\right)\right)$, $\tilde{f}_0(\cdot)$ being a spectral density, and $\tilde{\epsilon}(t)$ is a $K$-dimensional Gaussian white noise independent of $\alpha(0)$ with

$$\text{Cov}(\tilde{\epsilon}(t), \tilde{\epsilon}(t')) = \delta_{t,t'} \text{diag}\left(\tilde{f}(k_j)\right), \quad (3.11)$$

where $\tilde{f}(\cdot)$ is a spectral density. Then the process $\xi^K(t, s) = \phi(s)'\alpha(t)$, where the components $\alpha_j(t)$ are given by

$$\alpha_j(t) = \exp(h_j t) \alpha_j(0) + \int_0^t \exp(h_j(t - u)) \tilde{\epsilon}_j(u) du \quad (3.12)$$

with $h_j = -i \cdot \mu' k_j - k_j' \Sigma k_j - \zeta$, is a solution of the SPDE in (3.1). For $t \to \infty$, the influence of the initial condition $\exp(h_j t) \alpha_j(0)$ converges to zero and the process $\xi(t, s)$ converges to a stationary Gaussian process with mean zero and

$$\text{Cov}(\xi^K(t + \Delta t, s), \xi^K(t, s')) = \phi(s)' \text{diag}\left(-\exp(h_j \Delta t)\tilde{f}(k_j)\right) \phi(s')^*,$$

where $\cdot^*$ stands for complex conjugation.

This result shows that the solution of the SPDE is exact over time, given the frequencies included. In contrast to finite differences, one does not accumulate errors over time. The approximation error only depends on the number of spectral terms and not on the temporal discretization, see also Proposition 2 below. This is related to the fact that there is no need for numerical stability conditions. For statistical applications, where the parameters are not known a priori, this is particularly useful. Since Fourier terms are global functions, stationarity in space but not in time is a necessary assumption.

**Proof.** By (3.12), we have

$$\frac{\partial}{\partial t} \xi^K(t, s) = \sum_{j=1}^K \dot{\alpha}_j(t) \phi_j(s) = \sum_{j=1}^K (h_j \alpha_j(t) + \tilde{\epsilon}_j(t)) \phi_j(s).$$
On the other hand, since the functions $\phi_j(s) = \exp(i \cdot k_j' s)$ are Fourier terms, differentiation in the physical space corresponds to multiplication in the spectral space:

$$\mu' \nabla \phi_j(s) = i \mu' k_j \phi_j(s)$$ (3.13)

and

$$\nabla \cdot \Sigma \nabla \phi_j(s) = -k_j' \Sigma k_j \phi_j(s).$$ (3.14)

Therefore, by the definition of $h_j$,

$$(-\mu' \nabla + \nabla \cdot \Sigma \nabla - \zeta) \sum_{j=1}^{K} \alpha_j(t) \phi_j(s) = \sum_{j=1}^{K} h_j \alpha_j(t) \phi_j(s).$$

Together, we have

$$\frac{\partial}{\partial t} \xi^K(t, s) = (-\mu' \nabla + \nabla \cdot \Sigma \nabla - \zeta) \xi^K(t, s) + \epsilon^K(t, s)$$

which proves the first part of the proposition. Since the real part of $h_j$ is negative, $\exp(h_j t) \to 0$ for $t \to \infty$. Moreover,

$$\lim_{t \to \infty} \text{Cov}(\alpha_j(t + \Delta t), \alpha_{j'}(t))$$

$$= \lim_{t \to \infty} \exp(h_j \Delta t) \delta_{j,j'} \tilde{f}(k_j) \int_0^t \exp(-(h_j + h_{j'}^*)(t - u)) \, du$$

$$= -\frac{\exp(h_j \Delta t)}{h_j + h_{j'}^*} \delta_{j,j'} \tilde{f}(k_j),$$ (3.15)

and thus the last statement follows.

We assume that the forcing term $\epsilon(t, .)$, the initial state $\xi(0, .)$, and consequently also the solution $\xi(t, .)$, are stationary in space. Recall the Cramér representation for a stationary field $\epsilon(t, .)$

$$\epsilon(t, s) = \int \exp(i \cdot k' s) \bar{c}_t(k)$$
Chapter 3. SPDE based modeling of large space-time data

$\tilde{\epsilon}_t$ has orthogonal increments $\text{Cov}(d\tilde{\epsilon}_t(k), d\tilde{\epsilon}_{t'}(l)) = \delta_{t,t'} \delta_{k,l} \tilde{f}(k)$ and $\tilde{f}(\cdot)$ is the spectral density of $\epsilon(t,.)$ (see, e.g., Cramér and Leadbetter (1967)). This implies that we can approximate any stationary field, in particular also the one with a Whittle covariance function, by a finite linear combination of complex exponentials, and the covariance of $\tilde{\epsilon}(t)$ is a diagonal matrix as required in the proposition. Its entries are specified in (3.6). Concerning the initial state, one can use the stationary distribution of $\xi(t,.)$. An alternative choice is to use the same spatial distribution as for the innovations: $\tilde{f}_0(\cdot) = \tilde{f}(\cdot)$.

3.3.1 Approximation bound

By passing to the limit $K \to \infty$ such that both the wavenumbers $k_j$ cover the entire domain $\mathbb{R}^2$ and the distance between neighboring wavenumbers goes to zero, we obtain from (3.8) the stationary (in space and time) solution with spectral density as in (3.3). In practice, if one uses the discrete Fourier transform (DFT), or its fast variant, the FFT, the wavenumbers are regularly spaced and the distance between them is fixed for all $K$ (see below). This implies that the covariance function of an approximate solution is periodic which is equivalent to assuming a rectangular domain being wrapped around a torus. Since in most applications, the domain is fixed anyway, this is a reasonable assumption.

Based on the above considerations, we assume, in the following, that $s \in [0,1]^2$ with periodic boundary condition, i.e., that $[0,1]^2$ is wrapped on a torus. In practice, to avoid spurious periodicity, we can apply what is called “padding”. This means that we take $s \in [0,0.5]^2$ and then embed it in $[0,1]^2$. As in the discrete Fourier transform, if we choose $s \in [0,1]^2$, it follows that the spatial wavenumbers $k_j$ lie on the $n \times n$ grid given by $D_n = \{2\pi \cdot (i,j) : -(n/2 + 1) \leq i,j \leq n/2\} = \{-2\pi(n/2+1), \ldots, 2\pi n/2\}^2$ with $n^2 = N = K$, $n$ being an even natural number. We then have the following convergence result.

**Proposition 2.** The approximation $\xi^N(t,s)$ converges in law to the solution $\xi(t,s)$ of the SPDE (3.1) with $s \in [0,1]^2$ wrapped on a torus, and we have the bound

$$|C(t,s) - C^N(t,s)| \leq \sigma_\xi^2 - \sigma_{\xi^N}^2,$$

(3.16)
where \( C(t, s) \) and \( C^N(t, s) \) denote the covariance functions of \( \xi(t, s) \) and \( \xi^N(t, s) \), respectively, and where \( \sigma_\xi^2 = C(0, 0) \) and \( \sigma_{\xi^N}^2 = C^N(0, 0) \) denote the marginal variances of these two processes.

**Proof.** Similarly as in (3.4) and due to \( k \in 2\pi \cdot \mathbb{Z}^2 \), it follows that the covariance function of \( \xi(t, s) \) is given by

\[
C(t, s) = \sum_{k \in 2\pi \cdot \mathbb{Z}^2} \int f(\omega, k) \exp(i \cdot t \omega) d\omega \exp(i \cdot s' k),
\]

where \( h_k = -i \cdot \mu' k - k' \Sigma k - \zeta \). From Proposition 1 we know that the approximate solution \( \xi^N(t, s) \) has the covariance function

\[
C^N(t, s) = \sum_{k \in D_n} \tilde{f}(k) \frac{-\exp(h_k t)}{h_k + h_k^*} \exp(i \cdot s' k),
\]

It follows that

\[
|C(t, s) - C^N(t, s)| = \left| \sum_{k \in 2\pi \cdot \mathbb{Z}^2} \tilde{f}(k) \frac{-\exp(h_k t)}{h_k + h_k^*} (1 - \mathbbm{1}_{\{k \in D_n\}}) \exp(i \cdot s' k) \right|
\]

\[
\leq \sum_{k \in 2\pi \cdot \mathbb{Z}^2} \tilde{f}(k) \frac{1}{h_k + h_k^*} (1 - \mathbbm{1}_{\{k \in D_n\}})
\]

\[
= \sigma_\xi^2 - \sigma_{\xi^N}^2.
\]

(3.19)

Not surprisingly, this result tells us that the rate of convergence essentially depends on the smoothness properties of the process \( \xi(t, s) \), i.e., on how fast the spectrum decays. The smoother \( \xi(t, s) \), that is, the more variation is explained by low frequencies, the faster is the convergence of the approximation.

Note that there is a conceptual difference between the stationary solution of the SPDE (3.1) with \( s \in \mathbb{R}^2 \) and the periodic one with
\( s \in [0, 1]^2 \) wrapped on a torus. For the sake of notational simplicity, we have denoted both of them by \( \xi(t, s) \). The finite dimensional solution \( \xi^N(t, s) \) is an approximation to both of the above infinite dimensional solutions. The above convergence result, though, only holds true for the solution on the torus.

### 3.3.2 Real Fourier Functions and Discretization in Time and Space

In order that we can apply the model to real data, we have to discretize it. In the following, we consider the process \( \xi(t, s) \) on a regular grid of \( n \times n = N \) spatial locations \( s_1, \ldots, s_N \) in \([0, 1]^2\) and at equidistant time points \( t_1, \ldots, t_T \) with \( t_i - t_{i-1} = \Delta \). Note that these two assumptions can be easily relaxed, i.e., one can have irregular spatial observation locations and non-equidistant time points. The former can be achieved by adopting a data augmentation approach (see, for instance, Sigrist et al. (2012)) or by using an incidence matrix (see Section 3.4.2). The latter can be done by taking a time varying \( \Delta \).

For the sake of illustration, we have stated the results in the previous section using complex Fourier functions. However, when discretizing the model, one obtains a linear Gaussian state space model with a propagator matrix \( G \) that contains complex numbers, due to (3.13). To avoid this, we replace the complex terms \( \exp (i \cdot k_j' s) \) with real \( \cos(k_j' s) \) and \( \sin(k_j' s) \) functions. In other words, we use the real instead of the complex Fourier transform. The above results then still hold true, since for real valued data, the real Fourier transform is equivalent to the complex one. For notational simplicity, we will drop the superscript \( "K" \) from \( \xi^K(t, s) \). The distinction between the approximation and the true solution is clear from the context.

**Proposition 3.** On the above specified discretized spatial and temporal domain and using the real Fourier transform, a stationary solution of the SPDE (3.1) is of the form

\[
\begin{align*}
\xi(t_{i+1}) &= \Phi \alpha(t_{i+1}), \quad (3.20) \\
\alpha(t_{i+1}) &= G\alpha(t_i) + \tilde{\epsilon}(t_{i+1}), \quad \tilde{\epsilon}(t_{i+1}) \sim N(0, \tilde{Q}), \quad (3.21)
\end{align*}
\]
with stacked vectors $\xi(t_i) = (\xi(t_i, s_1), \ldots, \xi(t_i, s_N))'$ and cosine and sine coefficients $\alpha(t_i) = \left(\alpha_1^{(c)}(t_i), \ldots, \alpha_4^{(c)}(t_i), \alpha_5^{(c)}(t_i), \alpha_5^{(s)}(t_i), \ldots, \alpha_{K/2+2}^{(c)}(t_i), \alpha_{K/2+2}^{(s)}(t_i)\right)'$, where $\Phi$ applies the discrete, real Fourier transformation, $G$ is a block diagonal matrix with $2 \times 2$ blocks, and $\tilde{Q}$ is a diagonal matrix. These three matrices are defined as follows.

- $\Phi = [\phi(s_1), \ldots, \phi(s_N)]'$, where

$$
\phi(s_l) = \left(\phi_1^{(c)}(s_l), \ldots, \phi_4^{(c)}(s_l), \phi_5^{(c)}(s_l), \phi_5^{(s)}(s_l), \ldots, \phi_{K/2+2}^{(c)}(s_l), \phi_{K/2+2}^{(s)}(s_l)\right)',
$$

- $\phi_j^{(c)}(s_l) = \cos(k'_j s_l)$, $\phi_j^{(s)}(s_l) = \sin(k'_j s_l)$, $l = 1, \ldots, n^2$

- $[G]_{1:4,1:4} = \text{diag} \left(\exp \left(-\Delta (k'_j \Sigma k_j + \zeta)\right)\right)$, $[G]_{5:K,5:K} = \text{diag} \left(\exp \left(-\Delta (k'_j \Sigma k_j + \zeta)\right) \cdot \left(\cos(\Delta \mu' k_j) \mathbf{1}_2 - \sin(\Delta \mu' k_j) \mathbf{J}_2\right)\right)$, where

$$
\mathbf{1}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{J}_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad (3.22)
$$

- $\tilde{Q} = \text{diag} \left(\frac{1}{2} \left(\exp \left(-2 \Delta (k'_j \Sigma k_j + \zeta)\right) / (k'_j \Sigma k_j + \zeta)\right)\right)$.

In summary, at each time point $t$ and spatial point $s_l$, $l = 1, \ldots, n^2$, the solution $\xi(t, s_l)$ is the discrete real Fourier transform of the random coefficients $\alpha(t)$

$$
\xi(t, s_l) = \sum_{j=1}^{4} \alpha_j^{(c)}(t) \phi_j^{(c)}(s_l) + \sum_{j=5}^{K/2+2} \left(\alpha_j^{(c)}(t) \phi_j^{(c)}(s_l) + \alpha_j^{(s)}(t) \phi_j^{(s)}(s_l)\right) = \phi(s_l)' \alpha(t), \quad (3.23)
$$
and the Fourier coefficients \( \alpha(t) \) evolve dynamically over time according to the vector autoregression in (3.21). The first four terms are cosine terms and, afterwards, there are cosine - sine pairs. This is a peculiarity of the real Fourier transform. It is due to the fact that for four wavenumbers \( k_j \), the sine terms equal zero on the grid, i.e., \( \sin(k_j^l s_l) = 0 \), for all \( l = 1, \ldots, n^2 \) and \( k_j \in \{(0, 0)', (0, n\pi)', (n\pi, 0)', (n\pi, n\pi)\}' \) (see Figure 3.3). The above equations (3.20) and (3.21) form a linear Gaussian state space model with parametric propagator matrix \( G \) and innovation covariance matrix \( \tilde{Q} \), the parametrization being determined by the corresponding SPDE.

The model in (3.20) and (3.21) is similar to the one discussed Cressie and Wikle (2011, Chapter 7), but the derivation as an exact solution to the stochastic PDE (3.1) rather than a deterministic PDE is different.

Proof. Similarly as in Proposition 1, we first derive the continuous time solution. Using

\[
\mu' \nabla \phi_j^{(c)}(s_l) = -\mu' k_j \phi_j^{(s)}(s_l), \quad \mu' \nabla \phi_j^{(s)}(s_l) = \mu' k_j \phi_j^{(c)}(s_l),
\]

\[
\nabla \cdot \Sigma \nabla \phi_j^{(c)}(s_l) = -k_j' \Sigma k_j \phi_j^{(c)}(s_l), \quad \nabla \cdot \Sigma \nabla \phi_j^{(s)}(s_l) = -k_j' \Sigma k_j \phi_j^{(s)}(s_l),
\]

and the same arguments as in the proof of Proposition 1, it follows that the continuous time solution is of the form (3.23). For each pair of cosine - sine coefficients \( \alpha_j(t) = (\alpha_j^{(c)}(t), \alpha_j^{(s)}(t))' \) we have

\[
\alpha_j(t) = e^{H_j t} \alpha_j(0) + \int_0^t e^{H_j (t-u)} \tilde{\epsilon}_j(u) du, \quad (3.24)
\]

where

\[
H_j = \begin{pmatrix}
-k_j' \Sigma k_j - \zeta & -\mu' k_j \\
\mu' k_j & -k_j' \Sigma k_j - \zeta
\end{pmatrix}.
\]

Now \( H_j \) can be written as

\[
H_j = (-k_j' \Sigma k_j - \zeta) 1_2 - \mu' k_j J_2,
\]

where

\[
1_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad J_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.
\]
Since $1_2$ and $J_2$ commute, we have
\begin{align*}
 e^{H_j t} &= \exp \left( -t(k'_j \Sigma k_j + \zeta)1_2 \right) \exp \left( -t\mu' k_j J_2 \right) \\
 &= \exp \left( -t(k'_j \Sigma k_j + \zeta) \right) (\cos(t\mu' k_j)1_2 - \sin(t\mu' k_j)J_2).
\end{align*}

(3.25)

For the calculation of the exponential function of the matrix $J_2$, see, e.g., Bronson and Costa (2007, Chapter 4).

Analogously, one derives for the first four cosine terms
\begin{align*}
\alpha^c_j(t) &= e^{-(k'_j \Sigma k_j + \zeta)t} \alpha^c_j(0) + \int_0^t e^{-(k'_j \Sigma k_j + \zeta)(t-u)} \bar{\epsilon}_j(u) du, \quad j = 1, \ldots, 4.
\end{align*}

(3.26)

The above expression (3.25) and (3.26) give the propagator matrix $G$.

For the discrete time solution, in addition to the propagation
\[ \alpha_j(t + \Delta) = e^{H_j \Delta} \alpha_j(t), \]
we need to calculate the covariance of the integrated stochastic innovation term
\[ \int_t^{t+\Delta} e^{H_j(t+\Delta-u)} \bar{\epsilon}_j(u) du. \]

This is calculated as
\begin{align*}
\int_t^{t+\Delta} e^{H_j(t+\Delta-u)} \bar{f}(k_j) e^{H'_j(t+\Delta-u)} du \\
&= \int_0^\Delta e^{H_j(\Delta-u)} \bar{f}(k_j) e^{H'_j(\Delta-u)} du \\
&= \int_0^\Delta \bar{f}(k_j) \exp \left( -2(k'_j \Sigma k_j + \zeta)(\Delta - u) \right) 1_2 du \\
&= \bar{f}(k_j) \frac{1 - \exp \left( -2(k'_j \Sigma k_j + \zeta)\Delta \right)}{2(k'_j \Sigma k_j + \zeta)} 1_2.
\end{align*}

For the first four cosine term, calculations are done analogously. This then gives the result in (3.20) and (3.21).

The discrete complex Fourier transform uses $n^2$ different wavenumbers $k_j$ each having a corresponding Fourier term $\exp(i \cdot k'_j s)$. The real
Fourier transform, on the other hand, uses $n^2/2 + 2$ different wavenumbers, where four of them have only a cosine term and the others each have sine and cosine terms. This follows from the fact that, for real data, certain coefficients of the complex transform are the complex transpose of other coefficients. For technical details on the real Fourier transform, we refer to Dudgeon and Mersereau (1984), Borgman et al. (1984), Royle and Wikle (2005), and Paciorek (2007). Figure 3.2 illustrates an example of the spatial wavenumbers, with $n^2 = 20 \times 20 = 400$ grid points. The dots with a circle represent the wavenumbers actually used in the real Fourier transform, and the red crosses mark the wavenumbers having only a cosine term. Note that in (3.23) we choose to order the spatial wavenumbers such that the first four spatial wavenumbers correspond to the cosine-only terms. To get an idea how the basis functions $\cos(k'_j s)$ and $\sin(k'_j s)$ look like, we plot in Figure 3.3 twelve low-frequency basis functions corresponding to the six spatial frequencies closest to the origin $0$. Further, in Figure 3.4, there is an example of a propagator matrix $G$ when $n = 4$, i.e., when sixteen ($4^2$) spatial basis functions are used. The upper left $4 \times 4$ diagonal elements

\begin{figure}
\centering
\includegraphics[width=\textwidth]{spatial_wavenumbers.png}
\caption{Illustration of spatial wavenumbers for the two-dimensional discrete real Fourier transform with $n^2 = 400$ grid points.}
\end{figure}
3.3. Solution in the Spectral Space

Figure 3.3: Illustration of two dimensional Fourier basis functions used in the discrete real Fourier transform with $n^2 = 400$. On the $x$- and $y$-axis are the coordinates of $s$.

Figure 3.4: Illustration of propagator matrix $G$. 16 real Fourier functions are used ($n = 4$).

matrix corresponds to the cosine-only frequencies. The $2 \times 2$ blocks following correspond to wavenumbers with cosine - sine pairs.

Concerning notation in this paper, $K$ refers to the number of Fourier terms, i.e., this is the dimension of the spectral process $\alpha(t)$ at each time $t$. Furthermore, $N$ denotes the number of points at which the process $\xi(t)$ is modeled, and $n$ is the number of points on each axis of
the quadratic grid used. Often, we have \( n^2 = N = K \). However, if one uses a reduced dimensional Fourier basis, \( K \) is smaller than \( N \), see Section 3.4.2.

### 3.3.3 Remarks on Finite Differences

Another approach to solve PDEs or SPDEs such as the one in (3.1) consists of using a discretization such as finite differences. Stroud et al. (2010) use finite differences to solve an advection-diffusion PDE. Other examples are Wikle (2003), Xu and Wikle (2007), Duan et al. (2009), Malmberg et al. (2008), and Zheng and Aukema (2010). The finite difference approximation, however, has several disadvantages. First, each spatial discretization effectively implies an interaction structure between temporal and spatial correlation. In other words, as Xu et al. (2005) state, the discretization effectively suggests a knowledge of the scale of interaction, lagged in time. Usually, this space-time covariance interaction structure is not known, though. Furthermore, there are numerical stability conditions that need to be fulfilled in order that the approximate solution is meaningful. Since these conditions depend on the values of the unknown parameters, one can run into problems.

In addition, computational tractability is an issue. In fact, we have tried to solve the SPDE in (3.1) using finite differences as described in the following. A finite difference approximations in (3.1) leads to a vector autoregressive model with a sparse propagator matrix being determined by the discretization. The innovation term \( \epsilon \) can be approximated using a Gaussian Markov random field with sparse precision matrix (see Lindgren et al. (2011)). Even though the propagator and the precision matrices of the innovations are sparse, we have run into a computational bottleneck when using the Forward Filtering Backward Sampling (FFBS) algorithm (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994) for fitting the model. The basic problem is that the Kalman gain is eventually a dense matrix. Alternative sampling schemes like the information filter (see, e.g., Anderson and Moore (1979) and Vivar and Ferreira (2009)) did not solve the problem either. However, future research on this topic might come up with solutions.
3.4 Computationally Efficient Statistical Inference

As said in the introduction, when using a naive approach, the computational costs for a spatio-temporal model with $T$ time points and $N$ spatial points equal $O((NT)^3)$. Using the Kalman filter or the Forward Filtering Backward Sampling (FFBS) algorithm (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994), depending on what is needed, these costs are reduced to $O(TN^3)$ which, generally, is still too high for large data sets. In the following, we show how Bayesian and frequentist inference can be done efficiently in $O(TN \log N)$ operations. In the spectral space, the costs of the algorithms grow linearly in the dimension $TN$, which means that the total computational costs are dominated by the costs of the fast Fourier transform (FFT) (Cooley and Tukey, 1965) which are $O(TN \log N)$. Furthermore, computational time can be reduced by running the $T$ different FFTs in parallel.

As is often done in a statistical model, we add a non-structured Gaussian term $\nu(t_{i+1}, s) \sim N(0, \tau^2), \text{iid}$, to (3.20) to account for small scale variation and/or measurement errors. In geostatistics, this term is called nugget effect. Denoting the observations at time $t_i$ by $w(t_i)$, we then have the following linear Gaussian state space model:

\[
\begin{align*}
    w(t_{i+1}) &= \Phi \alpha(t_{i+1}) + \nu(t_{i+1}), \\
    \nu(t_{i+1}) &\sim N(0, \tau^2 \mathbf{1}_N), \\
    \alpha(t_{i+1}) &= G \alpha(t_i) + \tilde{\epsilon}(t_{i+1}), \\
    \tilde{\epsilon}(t_{i+1}) &\sim N(0, \tilde{Q}).
\end{align*}
\]

Note that $\xi(t_{i+1}) = \Phi \alpha(t_{i+1})$. As mentioned before, irregular spatial data can be modeled by adopting a data augmentation approach (see Sigrist et al. (2012)) or by using an incidence matrix (see Section 3.4.2). For the sake of simplicity, a zero mean was assumed. Extending the model by including covariates in a regression term is straightforward. Furthermore, we assume normality. The model can be easily generalized to allow for data not following a Gaussian distribution. For instance, this can be done by including it in a Bayesian hierarchical model (BHM) (Wikle et al., 1998) and specifying a non-Gaussian distribution for $w|\xi$. The likelihood can then no longer be evaluated exactly. But approximate posterior probabilities can still be computed using, for instance, simulation based methods such as Markov chain Monte Carlo (MCMC) (see, e.g., Gilks et al. (1996) or Robert and Casella (2004)).
An additional advantage of BHMs is that these models can be extended, for instance, to account for temporal non-stationarity by letting one or several parameters vary over time.

3.4.1 Kalman Filtering and Backward Sampling in the Spectral Space

In a frequentist context, it is crucial that one is able to evaluate the likelihood with a reasonable computational effort, and when doing Bayesian inference, one needs to be able to simulate efficiently from the full conditional of the latent process \( \xi \), or, equivalently, the Fourier coefficients \( \alpha \). Below, we show how both these tasks can be done in the spectral space in linear time, i.e., using \( O(TN) \) operations. For transforming between the physical and spectral space, one can use the FFT which requires \( O(TN \log N) \) operations. We start with the spectral version of the Kalman filter. Its output is used for both evaluating the log-likelihood and for simulating from the full conditional of the coefficients \( \alpha \).

Algorithm 1 Spectral Kalman filter

**Input:** \( T, \tilde{w}, G, \tau^2, \tilde{Q}, F \)

**Output:** forecast and filter means \( m_{t_i \mid t_{i-1}},R_{t_i \mid t_{i-1}}, \) \( i = 1, \ldots, T \)

\[
\begin{align*}
m_{t_0 \mid t_0} & = 0 \\
R_{t_0 \mid t_0} & = \tilde{Q} \\
\text{for } i = 1 \text{ to } T \text{ do} \\
\quad m_{t_i \mid t_{i-1}} & = G m_{t_{i-1} \mid t_{i-1}} \\
\quad R_{t_{i-1} \mid t_{i-1}} & = \tilde{Q} + R_{t_{i-1} \mid t_{i-1}}F \\
\quad R_{t_i \mid t_i} & = \left( \tau^{-2}1_N + R_{t_{i-1} \mid t_{i-1}}^{-1} \right)^{-1} \\
\quad m_{t_i \mid t_i} & = m_{t_i \mid t_{i-1}} + \tau^{-2}R_{t_i \mid t_i} \left( \tilde{w}(t_i) - m_{t_i \mid t_{i-1}} \right) \\
\text{end for}
\end{align*}
\]

Algorithm 1 shows the Kalman filter in the spectral space. For the sake of simplicity, we assume that the initial distribution equals the innovation distribution. The spectral Kalman filter has as input
3.4. Computationally Efficient Statistical Inference

the Fourier transform of \( \tilde{w} = (\tilde{w}(t_1)', \ldots, \tilde{w}(t_T)')' \) of \( w \), the diagonal matrix \( F \) given by

\[
[F]_{1:4,1:4} = \text{diag} \left( \exp \left( -2\Delta (k_j' \Sigma k_j + \zeta) \right) \right), \\
[F]_{5:N,5:N} = \text{diag} \left( \exp \left( -2\Delta (k_j' \Sigma k_j + \zeta) \right) 1_2 \right),
\]

and other parameters that characterize the SPDE model. It returns forecast and filter means \( m_{t_i|t_{i-1}} \) and \( m_{t_i|t_i} \) and covariance matrices \( R_{t_i|t_i} \) and \( R_{t_i|t_{i-1}} \), \( i = 1, \ldots, T \), respectively. I.e., \( m_{t_i|t_i} \) and \( R_{t_i|t_i} \) are the mean and the covariance matrix of \( \alpha(t_i) \) given data up to time \( t_i \) \{\( w(t_j)|j = 1, \ldots, i \}\}. Analogously, \( m_{t_i|t_{i-1}} \) and \( R_{t_i|t_{i-1}} \) are the forecast mean and covariance matrix given data up to time \( t_{i-1} \). We follow the notation of Künsch (2001).

Since the matrices \( \tilde{Q} \) and \( F \) are diagonal, the covariance matrices \( R_{t_i|t_i} \) and \( R_{t_i|t_{i-1}} \) are also diagonal. Note that the matrix notation in Algorithm 1 is used solely for illustrational purpose. In practice, matrix vector products \( (G_m t_{i-1}|t_{i-1}) \), matrix multiplications \( (R_{t_i|t_{i-1}} 1F) \), and matrix inversions \( ((\tau^{-2} + R_{t_i|t_{i-1}})^{-1}) \) are not calculated with general purpose algorithms but elementwise since all matrices are diagonal. It follows that the computational costs for this algorithm are \( O(TN) \).

The derivation of this algorithm follows from the classical Kalman filter (see, e.g., Künsch (2001)) using \( \Phi^' \Phi = 1_N \), \( GR_{t_{i-1}|t_{i-1}} G' = R_{t_{i-1}|t_{i-1}} GG' \), and the fact that \( GG' = F \). The first equation holds true due to the orthonormality of the discrete Fourier transform. The second equation follows from the fact that \( G \) is \( 2 \times 2 \) block diagonal and that \( R_{t_{i-1}|t_{i-1}} \) is diagonal with the diagonal entries being equal for each cosine - sine pair. The last equation holds true as shown in the following. Being obvious for the first four frequencies, we consider the \( 2 \times 2 \) diagonal blocks of cosine - sine pairs:

\[
[G]_{(2l-5):(2l-4),(2l-5):(2l-4)} [G']_{(2l-5):(2l-4),(2l-5):(2l-4)} \\
= \exp \left( -2\Delta (k_j' \Sigma k_j + \zeta) \right) (\cos(\Delta \mu' k_j) 1_2 - \sin(\Delta \mu' k_j) J_2) \\
\cdot (\cos(\Delta \mu' k_j) 1_2 - \sin(\Delta \mu' k_j) J_2)' \\
= \exp \left( -2\Delta (k_j' \Sigma k_j + \zeta) \right) (\cos(\Delta \mu' k_j)^2 + \sin(\Delta \mu' k_j)^2) 1_2, \\
l = 5, \ldots, N/2 + 2,
\]
which equals (3.28). In the last equation we have
used

\[ J'_2 = -J_2 \quad \text{and} \quad J'_2 = -1_2. \]

Based on the Kalman filter, the log-likelihood is calculated as (see, e.g., Shumway and Stoffer (2000))

\[
\ell = \sum_{i=1}^{T} \left( \tilde{w}(t_i) - m_{t_i|t_i-1} \right)' \left( R_{t_i|t_i-1} + \tau^2 1_N \right)^{-1} \left( \tilde{w}(t_i) - m_{t_i|t_i-1} \right) \\
+ \sum_{i=1}^{T} \log \left| R_{t_i|t_i-1} + \tau^2 1_N \right| + \frac{TN}{2} \log(2\pi).
\]

(3.29)

Since the forecast covariance matrices \( R_{t_i|t_i-1} \) are diagonal, calculation of their determinants and their inverses is trivial, and computational costs are again \( O(TN) \).

**Algorithm 2** Spectral backward sampling

**Input:** \( T, G, \tilde{Q}, F, m_{t_i|t_i-1}, m_{t_i|t_i}, R_{t_i|t_i}, R_{t_i|t_i-1}, i = 1, \ldots, T \)

**Output:** a sample \( \alpha^*(t_1), \ldots, \alpha^*(t_T) \) from \([\alpha|\cdot]\)

\[
\alpha^*(t_T) = m_{t_T|t_T} + \left( R_{t_T|t_T} \right)^{1/2} n_T, \quad n_T \sim N(0,1_N)
\]

for \( i = T-1 \) to 1 do

\[
\bar{m}_{t_i} = m_{t_i|t_i} + R_{t_i|t_i} R_{t_i|t_i-1}^{-1} G' \left( \alpha^*(t_{i+1}) - m_{t_i|t_i-1} \right)
\]

\[
\bar{R}_{t_i} = \left( \tilde{Q} F + R_{t_i-1|t_i-1}^{-1} \right)^{-1}
\]

\[
\alpha^*(t_i) = \bar{m}_{t_i} + \left( \bar{R}_{t_i} \right)^{1/2} n_i, \quad n_i \sim N(0,1_N)
\]

end for

As said, in a Bayesian context, the main difficulty consists in simulating from the full conditional of the latent coefficients \([\alpha|\cdot]\). After running the Kalman filter, this can be done with a backward sampling step. Together, these two algorithms are know as Forward Filtering Backward Sampling (FFBS) (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994). Again, backward sampling is computationally very efficient in the spectral space with costs being \( O(TN) \). Algorithm 2 shows the backward sampling algorithm in the spectral space. The matrices \( \bar{R}_{t_i} \) are diagonal which makes their Cholesky decomposition trivial.
3.4.2 Dimension Reduction and Missing or Non-Gridded Data

If desired, the total computational costs can be additionally alleviated by using a reduced dimensional Fourier basis with $K << N$, $N$ being the number of grid points. This means that one includes only certain frequencies, typically low ones. When the Fourier transform has been made, the spectral filtering and sampling algorithms then require $O(KT)$ operations. For using the FFT, the frequencies being excluded are just set to zero. Performing the FFT still requires $O(TN \log N)$ operations, though.

When the observed data does not lie on a grid or has missing data, there are two alternative approaches. First, one can use a data augmentation approach (Smith and Roberts, 1993) for the missing data. See Section 3.5.3 and, for more details, Sigrist et al. (2012). For irregularly spaced data, one can assign the data to a regular grid and treat the cells with no observations as missing data. FFT can then be applied to the augmented data, and the algorithms presented above can be used. Alternatively, as is the case in our application, one can include an incidence matrix $I$ that relates the process on the grid to the observation locations. Instead of (3.27), the model is then

$$w(t_{i+1}) = I\Phi \alpha(t_{i+1}) + \nu(t_{i+1}), \quad \nu(t_{i+1}) \sim N(0, \tau^2 1_N). \quad (3.30)$$

However, in the Kalman filter, the term $(I\Phi)'I\Phi$, used for calculating the filter covariance matrix $R_{t_i|t_i}$, is not a diagonal matrix anymore. From this follows that the Kalman filter does not diagonalize in the spectral space if one uses an incidence matrix $I$. Consequently, one has to use the traditional FFBS for which computational costs are $O(K^3T)$. This means that dimension reduction is required to make this approach computationally feasible.

3.4.3 An MCMC algorithm for Bayesian inference

Based on the algorithms presented above, there are different possible ways for doing statistical inference. For instance, if one adopts a frequentist paradigm, one can numerically maximize the log-likelihood in (3.29). In the following, we briefly present how Bayesian inference can be done using a Monte Carlo Markov Chain (MCMC) algorithm (see
Chapter 3. SPDE based modeling of large space-time data

Gilks et al., 1996; Robert and Casella, 2004; Brooks et al., 2011). This algorithm is implemented in the R package \texttt{spate} (Sigrist et al., 2012a) and used in the application in Section 3.5.

To complete the specification of a Bayesian model, prior distributions for the parameters \( \theta = (\rho_0, \sigma^2, \zeta, \rho_1, \gamma, \alpha, \mu_x, \mu_y, \tau^2)' \) have to be chosen. In general, this choice can depend on the specific application. We present choices for priors that are weakly uninformative. Based on Gelman (2006), we suggest to use improper priors for the \( \sigma^2 \) (marginal variance of the innovation) and \( \tau^2 \) (nugget effect variance) that are uniform on the standard deviation scale \( \sigma \) and \( \tau \), respectively. Further, the drift parameters \( \mu_x \) and \( \mu_y \) have uniform priors on \([-0.5, 0.5]\), \( \psi \) (direction of anisotropy) has a uniform prior on \([0, \pi/2]\), and \( \gamma \) (degree of anisotropy) has a uniform prior on the log scale of the interval \([0.1, 10]\). \( \gamma \) is restricted to \([0.1, 10]\) since stronger anisotropy does not seem reasonable. The range parameters of the innovations and the diffusion matrix \( \rho_0 \) and \( \rho_1 \), respectively, as well as the damping parameter \( \zeta \) are assigned improper, locally uniform priors on \( \mathbb{R}_+ \).

Our goal is then to simulate from the joint posterior of the unobservables \( [\theta, \alpha|w] \), where \( w \) denotes the set of all observations. Missing data can be accommodated for by using a data augmentation approach which results in an additional Gibbs step, see Section 3.5.3. Since the latent process \( \xi \) is the Fourier transform of the coefficients \( \alpha \), \( \xi(t_i) = \Phi \alpha(t_i) \), sampling from posterior of \( \alpha \) is, from a methodological point of view, equivalent to sampling from the one of \( \xi \). In the following, we use the notation "\([w|\cdot]\)" and "\( P[w|\cdot]\)" to denote conditional distributions and densities, respectively.

A straightforward approach would be to sample iteratively from the full conditionals of \( \theta \) and \( \alpha \). One could also further divide the latent process \( \alpha \) in blocks by iteratively sampling \( \alpha(t_i) \) at each time point. However \( \theta \) and \( \alpha \) can be strongly dependent which results in slow mixing. This problem is similar to the one observed when doing inference for diffusion models, see, e.g., Roberts and Stramer (2001) and Golightly and Wilkinson (2008). It is therefore recommendable to sample jointly from \( [\theta, \alpha|w] \) in a Metropolis-Hastings step.

Joint sampling from \( \theta \) and \( \alpha \) is done as follows. First, a proposal \((\theta^*, \alpha^*)\) is obtained by sampling \( \theta^* \) from a Gaussian distribution with
3.5. Postprocessing Precipitation Forecasts

the mean equaling the last value and an adaptively estimated proposal covariance matrix. Then, a sample $\alpha^*$ from $[\alpha|\theta^*, w]$ is obtained using the forward filtering backward sampling (FFBS) algorithm (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994). It can be shown that the acceptance ratio for the joint proposal is

$$\min \left( 1, \frac{P[\theta^*|w]}{P[\theta^{(i)}|w]} \right),$$

where $P[\theta|w]$ denotes the likelihood of $\theta$ given $w$, and where $\theta^*$ and $\theta^{(i)}$ denote the proposal and the last values, respectively. We see that this acceptance ratio does not depend on the latent process $\xi = \Phi \alpha$. Thus, the parameters $\theta$ are allowed to move faster in their parameter space. The value of the likelihood $P[\theta|w]$ is obtained as a side product of the Kalman filter in the FFBS.

For this random walk Metropolis step, we suggest to use an adaptive algorithm (Roberts and Rosenthal, 2009) meaning that the proposal covariance matrices for $\theta$ are successively estimated such that an optimal scaling is obtained with an acceptance rate between 0.2 and 0.3. See Roberts and Rosenthal (2001) for more information on optimal scaling for Metropolis-Hastings algorithms.

In addition, if the model includes a regression term (see the application in Section 3.5), the fixed effects can also be strongly dependent with the random effects $\xi$. This means that it is advisable that the coefficients $b \in \mathbb{R}^p$ of the potential covariates $x(t, s) \in \mathbb{R}^p$ are also sampled together with $\theta$ and $\alpha$. This can be done by slightly modifying the above algorithm. First, the regression coefficients $b^*$ are proposed jointly with $\theta^*$ in random walk Metropolis step. Then $\alpha^*$ is sampled from $[\alpha|\theta^*, b^*, w]$ analogously using the FFBS. Finally, in the acceptance ratio in (3.31), $P[\theta|w]$ now just has to be replaced by $P[\theta, b|w]$ which is also a side product of the Kalman filter.

3.5 Postprocessing Precipitation Forecasts

Nowadays, numerical weather prediction (NWP) models are capable of producing predictive fields at spatially and temporally high frequen-
cies. Statistical postprocessing serves two purposes. First, probabilistic predictions are obtained in cases where only deterministic ones are available. Further, even if “probabilistic” forecasts in form of ensembles (Palmer, 2002; Gneiting and Raftery, 2005) are available, they are typically not calibrated, i.e., they are often underdispersed (Hamill and Colucci, 1997). The goal of postprocessing is then to obtain calibrated and sharp predictive distributions (see Gneiting et al. (2007a) for a definition of calibration and sharpness). In the case of precipitation, the need for postprocessing is particularly strong, since, despite their importance, precipitation forecasts are still not as accurate as forecasts for other meteorological quantities (Applequist et al., 2002; Stensrud and Yussouf, 2007).

Several approaches for postprocessing precipitation forecasts have been proposed, including linear regression (Antolik, 2000), logistic regression (Hamill et al., 2004), quantile regression (Bremnes, 2004; Friederichs and Hense, 2007), hierarchical models based on prior climatic distribution (Krzysztofowicz and Maranzano, 2006), neural networks (Ramrez et al., 2005), and binning techniques (Yussouf and Stensrud, 2006). Sloughter et al. (2007) propose a two-stage model to postprocess precipitation forecasts. Berrocal et al. (2008) extended the model of Sloughter et al. (2007) by accounting for spatial correlation. Kleiber et al. (2011) present a similar model that includes ensemble predictions and accounts for spatial correlation.

Except for the last two references, spatial correlation is typically not modeled in postprocessing precipitation forecasts, and none of the aforementioned models explicitly accounts for spatio-temporal dependencies. However, for temporally and spatially highly resolved data, it is necessary to account for correlation in space and time. First, spatio-temporal correlation is important, for instance, for predicting precipitation accumulation over space and time with accurate estimates of precision. Further, it is likely that errors of NWP models exhibit structured behaviour over space and time. A spatio-temporal model can capture structured dynamics of this error term and extrapolate a spatial error over time.
3.5. Postprocessing Precipitation Forecasts

3.5.1 Data

![Locations of forecast gridpoints and observation stations](image)

**Figure 3.5:** Locations of grid points at which predictions are obtained (50×100 grid of small dots) and observations stations (bold dots). Both axis are in km using the Swiss coordinate system (CH1903).

The goal is to postprocess precipitation forecasts from an NWP model called COSMO-2, a high-resolution model with a grid spacing of 2.2 km that is run by MeteoSwiss as part of COonsortium for Small-scale MOdelling (COSMO) (see, e.g., Steppeler et al., 2003). The NWP model produces deterministic forecasts once a day starting at 0:00UTC. Predictions are made for eight consecutive time periods corresponding to 24 h ahead. In the following, let $y_F(t, s)$ denote the forecast of the rainfall sum from time $t - 1$ to $t$ at site $s$ made at 0:00UTC of the same day. We consider a rectangular region in northern Switzerland
shown in Figure 3.5. The grid at which predictions are made is of size $50 \times 100$. Precipitation is observed at 32 stations over northern Switzerland. Figure 3.5 also shows the locations of the observation stations. In the postprocessing model, the NWP forecasts are used as covariates in a regression term, see (3.33). We use data for three hourly rainfall amounts from the beginning of December 2008 till the end of March 2009. To illustrate the observed data, in Figure 3.6, observed precipitation at one station and the equally weighted areal average precipitation are plotted versus time. We will use the first three months containing 720 time points for fitting, and the last month is left aside for evaluation.

**Figure 3.6:** Precipitation (mm) versus time, for one station and averaged over all stations.

The NWP model forecasts are deterministic and ensembles are not available in our case. However, the extension to use an ensemble instead of just one member can be easily done. One can include all the
ensemble members in the regression part of the model. Or, in the case of exchangeable members, one can use the location and the spread of the ensemble.

3.5.2 Precipitation Model for Postprocessing

The model presented in the following is a Bayesian hierarchical model (BHM). It uses the SPDE based spatio-temporal Gaussian process $\xi(t, s)$ presented in Section 3.3 at the process level. At the data stage, a model adapted to the nature of precipitation is used. A characteristic feature of precipitation is that its distribution consists of a discrete component, indicating occurrence of precipitation, and a continuous one, determining the amount (see Figure 3.6). As a consequence, there are two basic statistical modeling approaches. The continuous and the discrete part are either modelled separately (Coe and Stern, 1982; Wilks, 1999) or together (Bell, 1987; Wilks, 1990; Bardossy and Plate, 1992; Hutchinson, 1995; Sansó and Guenni, 2004). See, e.g., Sigrist et al. (2012) for a more extensive overview of precipitation models and for further details on the data model used below. Originally, the approach presented in the following goes back to Tobin (1958) who analyzed household expenditure on durable goods. For modeling precipitation, Stidd (1973) took up this idea and modified it by including a power-transformation for the non-zero part so that the model can account for skewness.

It is assumed that rainfall $y(t, s)$ at time $t$ on site $s \in \mathbb{R}^2$ depends on a latent Gaussian variable $w(t, s)$ through

$$
    y(t, s) = 0, \quad \text{if } w(t, s) \leq 0,
    = w(t, s)^\lambda, \quad \text{if } w(t, s) > 0, \quad (3.32)
$$

where $\lambda > 0$. A power transformation is needed since precipitation amounts are skewed and do not follow a truncated normal distribution. The latent Gaussian process $w(t, s)$ is interpreted as a precipitation potential.

The mean of the Gaussian process $w(t, s)$ is assumed to depend linearly on spatio-temporal covariates $\mathbf{x}(t, s) \in \mathbb{R}^k$. As shown below, this mean term basically consists of the NWP forecasts. Variation that is
not explained by the linear term is modeled using the Gaussian process \( \xi(t, s) \) and the unstructured term \( \nu(t, s) \) for microscale variability and measurement errors. The spatio-temporal process \( \xi(t, s) \) has two functions. First, it captures systematic errors of the NWP in space and time and can extrapolate them over time. Second, it accounts for structured variability so that the postprocessed forecast is probabilistic and its distribution sharp and calibrated.

To be more specific concerning the covariates, similarly as in Berrocal et al. (2008), we include a transformed variable \( y_F(t, s)^{1/\tilde{\lambda}} \) and an indicator variable \( 1_{\{y_F(t, s) = 0\}} \) which equals 1 if \( y_F(t, s) = 0 \) and 0 otherwise. \( \tilde{\lambda} \) is determined by fitting the transformed Tobit model as in (3.32) to the marginal distribution of the rain data ignoring any spatio-temporal correlation. In doing so, we obtain \( \tilde{\lambda} \approx 1.4 \). \( y_F(t, s)^{1/\tilde{\lambda}} \) is centered around zero by subtracting its overall mean \( \bar{y}_F^{1/\tilde{\lambda}} \) in order to reduce posterior correlations. Thus, \( w(t, s) \) equals

\[
  w(t, s) = b_1 \left( y_F(t, s)^{1/\tilde{\lambda}} - \bar{y}_F^{1/\tilde{\lambda}} \right) + b_2 1_{\{y_F(t, s) = 0\}} + \xi(t, s) + \nu(t, s).
\]

An intercept is not included since the first Fourier term is constant in space. In our case, including an intercept term results in weak identifiability which slows down the convergence of the MCMC algorithm used for fitting. Note that in situations where the mean is large it is advisable to include an intercept, since the coefficient of the first Fourier term is constrained by the joint prior on \( \alpha \). Further, unidenstifiability is unlikely to be a problem in these cases.

Concerning the spatio-temporal process \( \xi(t, s) \), we apply padding. This means that we embed the 50 × 100 grid in a rectangular 200 × 200 grid. A brief prior investigation showed that the range parameters are relatively large in comparison to the spatial domain, and padding is therefore used in order to avoid spurious correlations due to periodicity. The NWP forecasts are not available on the extended 200 × 200 domain, which means that, in principle, the process \( w(t, s) \) can only be modeled on the 50 × 100 grid where the covariates are available. To cope with this we use an incidence matrix \( I \) as in (3.30) to relate the process at the 200 × 200 grid to the observation stations. As argued in Section 3.4.2, this then requires that we use a reduced dimensional Fourier
expansion. I.e., instead of using \( N = 200^2 \) basis functions, we only use \( K \ll N \) low-frequency Fourier terms. Since the observation stations are relatively scarce, one might argue that there is no information on spatial high frequencies of the NWP error, and that the high frequencies can be left out. In fact, this hypothesis gets confirmed by our analysis, see Figure 3.7.

Concerning prior distributions, for \( \theta = (\rho_0, \sigma^2, \zeta, \rho_1, \gamma, \psi, \mu_x, \mu_y, \tau^2)' \), we use the priors presented in Section 3.4.3. The parameters \( b \) and \( \lambda \), which are not included in \( \theta \), have improper, locally uniform priors on \( \mathbb{R} \) and \( \mathbb{R}_+ \), respectively. In summary,

\[
P[b, \lambda, \theta] \propto \frac{1}{\sqrt{\sigma^2 \tau^2 \gamma}} 1\{-0.5 \leq \mu_x, \mu_y \leq 0.5\} 1\{0 \leq \psi \leq \pi/2\} 1\{\lambda, \rho_0, \rho_1, \zeta, \sigma^2, \tau^2 \geq 0\} 1\{0.1 \leq \gamma \leq 10\}.
\]

In addition, concerning \( \alpha(0) \), we choose to use the innovation distribution specified in (3.6) as initial distribution.

### 3.5.3 Fitting

As mentioned before, Monte Carlo Markov Chain (MCMC) is used to sample from the posterior distribution \( [b, \lambda, \theta, \alpha, w|y] \), where \( y \) denotes the set of all observations. We use what Neal and Roberts (2006) call a Metropolis within-Gibbs algorithm which alternates between blocked Gibbs (Gelfand and Smith, 1990) and Metropolis (Metropolis et al., 1953; Hastings, 1970) sampling steps.

The main part of the algorithm consists of the MCMC algorithm presented in Section 3.4.3 with the coefficients \( b \) being sampled jointly with \( \theta \) and \( \alpha \). Due to the non-Gaussian data model, additional Metropolis and Gibbs steps are required for \( \lambda \) and for those points of \( w \) where the observed rainfall amount is zero and where observations are missing. We refer to Sigrist et al. (2012) for more details on the type of data augmentation approach that is used for doing this. We denote by \( w^{[0]} \) the values of \( w \) at those points where the observed rainfall is zero \( y(t,s) = 0 \). Analogously, we define \( w^{[m]} \) and \( w^{[+]} \) for the missing values \( y(t,s) = NA \) and the values where a positive rainfall amount is observed \( y(t,s) > 0 \). The full conditionals of the censored \( w^{[0]} \) and
missing points $w^{[m]}$ are truncated and regular one-dimensional Gaussian distributions, respectively. Sampling from them is done in Gibbs steps. The transformation parameter $\lambda$ is sampled using a random walk Metropolis step. If a new value is accepted, $w^{[+]}$ needs to be updated using the deterministic relation $w(t, s) = y(t, s)^{1/\lambda}$ due to (3.32). From these Gibbs and Metropolis steps, we obtain $w$ consisting of simulated and transformed observed data. In the second part of the algorithm, we sample $b$, $\theta$, and $\alpha$ jointly from $[b, \theta, \alpha|w]$ using the algorithm presented in Section 3.4.3, where $w$ acts as if it was the observed data. After a burn-in of 5,000 iterations, we use 100,000 samples from the Markov chain to characterize the posterior distribution. Convergence is monitored by inspecting trace plots.

### 3.5.4 Model Selection and Results

As said, we use a reduced dimensional approach. The number of Fourier functions is determined based on predictive performance for the 240 time points that were set aside. We start with models including only spatial low-frequencies and add successively higher frequencies. In doing so, we only consider models that have the same resolution in each direction. For instance, we do not consider models that have higher frequency spatial basis functions in the east-west direction than in the north-south one.

In order to assess the performance of the predictions and to choose the number of basis functions to include, we use the continuous ranked probability score (CRPS) (Matheson and Winkler, 1976). The CRPS is a strictly proper scoring rule (Gneiting and Raftery, 2007) that assigns a numerical value to probabilistic forecasts and assesses calibration and sharpness simultaneously (Gneiting et al., 2007a). It is defined as

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - 1_{\{y \leq x\}})^2 dx,$$

where $F$ is the predictive cumulative distribution, $y$ is the observed realization, and $1$ denotes an indicator function. If a sample $y^{(1)}, \ldots, y^{(m)}$
from \( F \) is available, it can be approximated by

\[
\frac{1}{m} \sum_{i=1}^{m} |y^{(i)} - y| - \frac{1}{2m^2} \sum_{i,j=1}^{m} |y^{(i)} - y^{(j)}|. \tag{3.35}
\]

Ideally, one would run the full MCMC algorithm at each time point \( t \geq 720 \), including all data up to the point, and obtain predictive distributions from this. Since this is rather time consuming, we make the following approximation. We assume that the posterior distribution of the “primary” parameters \( \theta, b, \) and \( \lambda \) given \( y_{1:t} = \{y_1, \ldots, y_t\} \) is the same for all \( t \geq 720 \). That is, we neglect the additional information that the observations in March give about the primary parameters. Thus, the posterior distributions of the primary parameters are calculated only once, namely on the data set from December 2008 to February 2009. This assumption that the posterior of the primary parameters does not change with additional data may be questionable over longer time periods and when one moves away from the time period from which data is used to obtain the posterior distribution. But since all our data lies in the winter season, we think that this assumption is reasonable. If longer time periods are considered, one could use sliding training windows or model the primary parameters as non-stationary using a temporal evolution.

For each time point \( t \geq 720 \), we make up to 8 steps ahead forecasts corresponding to 24 hours. I.e., we sample from the predictive distribution of \( y^*_{t+k}, k = 1, \ldots, 8 \), given \( y_{1:t} = \{y_1, \ldots, y_t\} \).

In Figure 3.7, the average CRPS of the pointwise predictions and the areal predictions are shown for the different statistical models. In the left plot, the mean is taken over all stations and lead times, whereas the areal version is an average over all lead times. This is done for the models with different numbers of basis functions used. Models including only a few low-frequency Fourier terms perform worse. Then the CRPS decreases successively. The model based on including \( K = 29 \) Fourier functions performs best. After this, adding higher frequencies results into lower predictive performance. We interpret this results in the way that the observation data does not allow for resolving high frequencies in the error term between the forecasted and observed precipitation. Note that high frequencies of the precipitation process itself
Figure 3.7: Comparison of different statistical models using the continuous ranked probability score (CRPS). On the left are CRPSs of station specific forecasts and on the right are CRPSs of areal forecasts. \( K \) denotes the number of basis functions used in the model. “Sep” denotes the separable model with \( K = 29 \) Fourier terms. The unit of the CRPS is mm.

are accounted for by the forecast \( y_F \). For comparison, we also fit a separable model which is obtained by setting \( \mu = 0 \) and \( \Sigma^{-1} = 0 \). Concerning the number of Fourier functions, we use \( K = 29 \) different Fourier terms. The separable model clearly performs worse than the model with a non-separable covariance structure. Based on these findings, we decide to use the model with 29 cosine and sine functions.

Table 3.1 shows posterior medians as well as 95% credible intervals for the different parameters. Note that the range parameters \( \rho_0 \) and \( \rho_1 \) as well as the drift parameters \( \mu_x \) and \( \mu_y \) have been transformed back from the unit \([0, 1]\) scale to the original km scale. The posterior median of the variance \( \sigma^2 \) of the innovations of the spatio-temporal process is around 0.8. Compared to this, the nugget variance being about 0.3 is smaller. For the innovation range parameter \( \rho_0 \), we obtain a value of about 25 km. And the range parameter \( \rho_1 \) that controls the
3.5. Postprocessing Precipitation Forecasts

Table 3.1: Posterior medians and 95\% credible intervals for the SPDE based spatio-temporal model presented in Section 3.3 with $K = 29$ Fourier terms.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Median</th>
<th>2.5%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_0$</td>
<td>25.4</td>
<td>18.8</td>
<td>32.4</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.838</td>
<td>0.727</td>
<td>0.994</td>
</tr>
<tr>
<td>$\zeta$</td>
<td>0.00655</td>
<td>0.000395</td>
<td>0.0156</td>
</tr>
<tr>
<td>$\rho_1$</td>
<td>48.8</td>
<td>42.1</td>
<td>57.1</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>4.33</td>
<td>3.34</td>
<td>6.01</td>
</tr>
<tr>
<td>$\psi$</td>
<td>0.557</td>
<td>0.49</td>
<td>0.617</td>
</tr>
<tr>
<td>$\mu_x$</td>
<td>6.73</td>
<td>0.688</td>
<td>12.9</td>
</tr>
<tr>
<td>$\mu_y$</td>
<td>-4.19</td>
<td>-8.55</td>
<td>-0.435</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>0.307</td>
<td>0.288</td>
<td>0.327</td>
</tr>
<tr>
<td>$y_F(t, s)^{1/\lambda}$</td>
<td>0.448</td>
<td>0.414</td>
<td>0.481</td>
</tr>
<tr>
<td>$1{y_F(t, s)=0}$</td>
<td>-0.422</td>
<td>-0.5</td>
<td>-0.344</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>1.67</td>
<td>1.64</td>
<td>1.7</td>
</tr>
</tbody>
</table>

amount of diffusion or, in other words, the amount of spatio-temporal interaction, is approximately 49 km. With $\gamma$ and $\psi$ being around 4 and 0.6, respectively, we observe anisotropy in the south-west to north-east direction. This is in line with the orography of the region, as the majority of the grid points lies between two mountain ranges: the Jura to the north-west and the Alps to the south-east. The drift points to the south-east, both parameters being rather small though. Further, the damping parameter $\zeta$ has a posterior median of about 0.01.

Next, we compare the performance of the postprocessed forecasts with the ones from the NWP model. In addition to the temporal cross-validation, we do the following cross-validation in space and time. We first remove six randomly selected stations from the data, fit the latent process to the remaining stations, and evaluate the forecasts at the stations left out. Concerning the primary parameters, i.e., all parameters except the latent process, we use the posterior obtained from the full data including all stations. This is done for computational simplicity and since this posterior is not very sensitive when excluding a few sta-
Chapter 3. SPDE based modeling of large space-time data

Table 3.2: Comparison of NWP model and statistically postprocessed forecasts (‘Stat PP’) using the mean absolute error (MAE). ‘Static’ denotes the constant forecast obtained by using the most recently observed data. The unit of the MAE is mm.

<table>
<thead>
<tr>
<th></th>
<th>Stat PP</th>
<th>NWP</th>
<th>Static</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stationwise</td>
<td>0.359</td>
<td>0.485</td>
<td>0.594</td>
</tr>
<tr>
<td>Areal</td>
<td>0.303</td>
<td>0.387</td>
<td>0.489</td>
</tr>
</tbody>
</table>

Since the NWP produces 8 step ahead predictions once a day, we only consider statistical forecasts starting at 0:00UTC. This is in contrast to the above comparison of the different statistical models for which 8 step ahead predictions were made at all time points and not just once for each day. We use the mean absolute error (MAE) for evaluating the NWP forecasts. In order to be consistent, we also generate point forecasts from the statistical predictive distributions by using medians, and then calculate the MAE for these point forecasts. In Table 3.2, the results are reported. For comparison, we also give the score for the static forecast that is obtained by using the most recently observed data. The postprocessed forecasts clearly perform better than the raw NWP forecasts. In addition, the postprocessed forecasts have the advantage that they provide probabilistic forecasts quantifying prediction uncertainty.

The statistical model produces a joint spatio-temporal predictive distribution that is spatially highly resolved. To illustrate the use of the model, we show several quantities in Figure 3.8. We consider the time point $t = 760$ and calculate predictive distributions over the next 24 hours. Predicted fields for the period $t = 761, \ldots, 768$ from the NWP are shown in the top left corner. On the right of it are pointwise medians obtained from the statistical forecasts. This is a period during which the NWP predicts too much rainfall compared to the observed data (results not shown). The figure shows how the statistical model corrects for this. For illustration, we also show one sample from the predictive distribution. To quantify prediction uncertainty, the difference between the third quartile and the median of the predictive distribution is plotted. These plots again show the growing uncertainty with
increasing lead time. Other quantities of interest (not shown here), that can be easily obtained, include probabilities of precipitation occurrence or various quantiles of the distribution.

3.6 Conclusion

We present a spatio-temporal model and corresponding efficient algorithms for doing statistical inference for large data sets. Instead of using the covariance function, we propose to use a Gaussian process defined through an SPDE. The SPDE is solved using Fourier functions, and we have given a bound on the precision of the approximate solution. In the spectral space, one can use computationally efficient statistical algorithms whose computational costs grow linearly with the dimension, the total computational costs being dominated by the fast Fourier transform. The space-time Gaussian process defined through the advection-diffusion SPDE has a nonseparable covariance structure and can be physically motivated. The model is applied to postprocessing of precipitation forecasts for northern Switzerland. The postprocessed forecasts clearly outperform the raw NWP predictions. In addition, they have the advantage that they quantify prediction uncertainty.

An interesting direction for further research would be to extend the SPDE based model to allow for spatial non-stationarity. For instance, the deformation method of Sampson and Guttorp (1992), where the process is assumed to be stationary in a transformed space and non-stationary in the original domain, might be a potential way. Furthermore, since the operators of the SPDE are local, one can define the SPDE on general manifolds and, in particular, on the sphere (see, e.g., Lindgren et al. (2011)). It would be interesting to investigate to which extent spectral methods can still be used in practice.
Figure 3.8: Illustration of postprocessed spatio-temporal precipitation fields for the period \( t = 761, \ldots, 768 \). The figure shows the NWP forecasts (a), pointwise medians of the predictive distribution (b), one sample from the predictive distribution (c), and the differences between the third quartile and the median of the predictive distribution (d). All quantities are in mm. Note that the scales are different in different figures.
With an increasing use of hierarchical Bayesian models in spatial and spatio-temporal statistics, amongst other fields, there is a need for model checking tools. We show how the mixed predictive distribution can be used for graphical model checking in a Bayesian context using simulation based methods. In contrast to the posterior predictive distribution, the mixed predictive distribution combines posterior and prior knowledge and thus avoids double use of data. Therefore, the mixed predictive distribution correctly detects model deficiencies in situations where the posterior predictive density fails to do so. This is illustrated with two simulated data sets and a real space-time data case study.

4.1 Introduction

Model diagnostics is an essential part of any statistical analysis, whichever method of inference is used (Box and Tiao, 1973). Conclusions based
on a model that is far from plausible can be very deceptive. Even if one adheres to the position that “essentially, all models are wrong, but some are useful” (Box and Draper, 1987), model criticism is important for guiding the search for models that are indeed useful. As Box and Jenkins (1970) state, model building is an iterative process during which alternative models are developed. Therefore, we need tools to detect model deficiencies which are not explicitly based on specific alternative models. See for instance O’Hagan (2003) for a more extensive discussion of model checking and model criticism.

Methods for model diagnostics typically compare features in the data with features that are implied by a model. This can be achieved either by graphical methods or by a discrepancy measure indicating incompatibility of the data with a model, usually accompanied by a measure of evidence against the model. For obtaining the latter, the considered model is often embedded as a sub-model within a larger model and then likelihood ratios or Bayes factors are computed (Jeffreys, 1961; Gelfand and Dey, 1994; Kass and Raftery, 1995). With graphical methods we can compare several features simultaneously and let the eye detect those where differences are largest. Which features one wants to look at depends on the goal of the specific application and the experience of the statistician. For instance, residual and Q-Q plots are commonly used for checking linear models. For geostatistical data, the empirical variogram for different lags is a natural summary statistic, or for point patterns, Ripley’s $K$-function.

In a Bayesian setting, model checking is routinely done using predictive distributions and comparing them with characteristics of observed data. Often these predictive distributions are easily calculated using simulation based methods. In the following, let $y$ denote the observed variables and $\theta$ the unobserved variables. In practice, probably the most widely used predictive distribution is the posterior predictive distribution (Rubin, 1984). Samples from the posterior predictive density are generated by simulating $\theta$ from its posterior and then, conditional on this, a new replicate $y^{new}$:

$$y^{new} \sim \int P[y|\theta]P[\theta|y^{obs}]d\theta,$$

where $y^{obs}$ denotes the observed data. Summary statistics $s_k(y^{obs})$
are then plotted against $E(s_k(y^{new}))$ or against a sample of values $(s_k(y^{new,i}); i = 1, \ldots, N)$. However, the posterior predictive density is not an adequate tool in the context of structured hierarchical models as explained in the following.

Let us assume that the parameters $\theta$ can be split into two parts $\xi$ and $\phi$. The former are latent variables whose dimension grows with the number of observations $y$, whereas $\phi$ denotes parameters whose dimension is independent of the dimension of $y$. We will call $\xi$ incidental variables and $\phi$ primary parameters. For instance, in a mixed effects model, $\xi$ are random effects, whereas $\phi$ contains the fixed effects, the hyperparameters of the random effects, and, potentially, a variance parameter of measurement errors. Or, in a spatial model, $\xi$ can be a Gaussian random field and $\phi$ hyperparameters for this field. Generally, in a Bayesian hierarchical model (Wikle et al., 1998), $\xi$ corresponds to the process model, and $\phi$ specifies its parameter model. The focus of model criticism then naturally lies on the process model which reflects prior scientific information, whereas the prior model on $\phi$ is typically as non-informative as possible.

The problem with the posterior predictive density stems from the fact that, in many cases, the prior on $\xi$ is rather flexible such that a posterior $\xi$ is strongly determined by $y^{obs}$ (Stern and Cressie, 2000; Marshall and Spiegelhalter, 2003). Consequently, when comparing $y^{new}$ with $y^{obs}$, it may seem as if the prior model on $\xi$ was adequate even though it is far from being so (see, e.g., Section 4.3.2 and, in particular, Figure 4.2). In other words, graphical methods based on samples from the posterior predictive density fail to show model deficiencies due to this double use of data: first for obtaining the posterior and then for deciding whether the posterior is compatible with the observed data. The examples in this paper illustrate this phenomenon.

An alternative to the posterior predictive density is the prior predictive distribution:

$$y^{new} \sim \int P[y | \theta] P[\theta] d\theta. \quad (4.2)$$

However, the marginal predictive distribution may reject a good model simply because the prior has too little mass for parameters which fit the data, or it may be so spread out that no discrepancy is detected.
As O’Hagan (2003) states, this approach has little diagnostic value in cases where prior information is weak.

Comparing the two predictive densities in (4.1) and (4.2), we see that the posterior predictive density uses the posterior for both the latent variables $\xi$ and the primary parameters $\phi$, whereas the prior predictive density uses the prior for both $\xi$ and $\phi$. Since the goal is to assess the prior process model for $\xi$, but not the prior on the (hyper-)parameters $\phi$, the posterior should be used for the latter, and the prior for $\xi$. This leads to the mixed predictive distribution (see (4.4) for a definition). Gelman et al. (1996) first stated that this approach can be useful with hierarchical models but did not further elaborate on it. Similarly, Bayarri and Berger (2000) criticize Bayesian p values based on the posterior predictive density. Instead they propose the partial posterior predictive p-value where the test statistic is distributed according to its prior rather than its posterior distribution, and posterior information is only included in aspects of the distribution that are not related to the test statistic.

Although the disadvantages of the posterior predictive density have been discussed at several instances in the literature (Stern and Cressie, 2000; Marshall and Spiegelhalter, 2003), its use is still widespread for diagnostics. The goal of this paper is two-fold. First, we provide examples where the mixed predictive distribution is successful in detecting serious model deficiencies, but the posterior predictive density fails. Second, we show how the mixed predictive distribution can be used for graphical model diagnostics, whereas earlier papers like Marshall and Spiegelhalter (2007) focused on p-values. In particular, we introduce new graphical tools which are suitable for diagnostic checking of complex spatial or spatio-temporal models. We believe that graphical methods are the preferred way for doing model criticism. Typically, it is hard or impossible to find a useful single discrepancy measure from which a p-value might be calculated, and graphs are much more powerful in capturing different potential model deficiencies.

Another approach to avoid double use of data is cross validation. Leave-one out cross validation can be implemented efficiently in a Bayesian framework using only samples from the posterior, see Gelfand and Dey (1994). However this is not an option for the type of applications that
we have in mind because it allows only for marginal comparison. For instance, if one is interested in the spatial dependency structure in a geostatistical model, marginal checking is clearly not sufficient. Cross-validation where large subsets of data are left out are more promising, but this is usually computationally very expensive. See also Gelfand et al. (1992) and Stern and Cressie (2000) for more information on cross-validation in the context of Bayesian model checking.

4.2 The mixed predictive density and its use

A Bayesian model consists of observable variables $y$ and unobservable variables $\theta = (\xi, \phi)$ which we split into a process part $\xi$ and a parameter part $\phi$ as described above. A model is defined by specifying the distribution $P[y|\xi, \phi]$ of the observables given the unobservables, called the likelihood, and the prior distribution $P[\xi, \phi] = P[\xi|\phi]P[\phi]$ of the unobservables.

As discussed in the introduction, when doing model checking, the fit of a model can be explored using the posterior predictive distribution (Rubin, 1984; Aitkin, 1991) and comparing characteristics of the posterior predictive density with characteristics of observations.

Definition 3 (Posterior predictive distribution). The posterior predictive distribution is the distribution of a new independent set of observables $y_{\text{new}}$ under the posterior model, the model with the unobservables $\phi$ and $\xi$ distributed according to their posterior distribution. It is defined as

$$y_{\text{new}} \sim P[y|y_{\text{obs}}] = \int P[y|\xi, \phi]P[\xi, \phi|y_{\text{obs}}]d\xi d\phi. \quad (4.3)$$

For hierarchical models, the posterior of the incidentals $P[\xi|\phi, y_{\text{obs}}]$ is typically strongly influenced by the data and this can mask model deficiencies. Such deficiencies can then lead to poor predictions of $y_{\text{pred}}$ at new points which do have related, but not the same incidental variables, as explained in the following. The predictive distribution for a
set of points $y^{\text{pred}}$ is calculated as

$$y^{\text{pred}} \sim P[y^{\text{pred}}|y^{\text{obs}}]$$

$$= \int P[y^{\text{pred}}|\xi^{\text{pred}}, \phi]P[\xi^{\text{pred}}|\xi, \phi]P[\xi, \phi|y^{\text{obs}}]d\xi d\xi^{\text{pred}} d\phi.$$

This shows that misspecifications in the prior $P[\xi|\phi]$ can have large consequences on the prediction density for the incidentals $P[\xi^{\text{pred}}|\xi, \phi]$ and thus on $P[y^{\text{pred}}|y^{\text{obs}}]$.

**Definition 4** (Mixed predictive distribution). The mixed predictive distribution is the distribution of a new independent set of observables $y^{\text{new}}$ under the model with $\phi$ distributed according to its posterior distribution and $\xi$ according to its prior distribution. It is defined as

$$y^{\text{new}} \sim \int P[y|\xi, \phi] \overbrace{P[\xi|\phi]}^{\text{prior}} \overbrace{P[\phi|y^{\text{obs}}}^{\text{posterior}}d\xi d\phi. \quad (4.4)$$

The term mixed is used due to the fact that for some unobservables (the parameters $\phi$) posterior knowledge is included whereas for others (the process variables $\xi$) prior information is used. The mixed predictive distribution is a marginalization of the joint posterior predictive distribution for $y$ and $\xi$:

$$(y^{\text{new}}, \xi^{\text{new}}) \sim \int P[y|\xi, \phi]P[\xi|\phi]P[\phi|y^{\text{obs}}]d\phi.$$

Simulating from the posterior predictive density and the mixed predictive distribution is typically easy, especially, when simulation based methods such as Monte Carlo Markov Chain are used for inference. For instance, the mixed predictive distribution can be approximated by

$$\int P[y|\xi, \phi]P[\xi|\phi]P[\phi|y^{\text{obs}}]d\xi d\phi \approx \frac{1}{m} \sum_{i=1}^{m} P[y|\xi^{(i)}, \phi^{(i)}], \quad (4.5)$$

where $\phi^{(i)}, i = 1, \ldots, m,$ are samples from the corresponding posterior density $P[\phi|y^{\text{obs}}]$ and $\xi^{(i)}, i = 1, \ldots, m$ are samples from $P[\xi|\phi^{(i)}]$. Simulation from the mixed predictive distribution can, therefore, be
done in three steps: Simulate $\phi^{(i)}$ from its posterior distribution $P[\phi|y^{obs}]$. Given $\phi^{(i)}$, simulate a $\xi^{(i)}$ from $P[\xi|\phi^{(i)}]$. Then given $\phi^{(i)}$ and $\xi^{(i)}$, simulate $y$ according to $P[y|\xi^{(i)}, \phi^{(i)}]$.

4.3 Two simulated examples of mixed predictive distribution based graphical model checking

4.3.1 A mixed effects model

In the following, we illustrate how the posterior predictive density can be deceptive in the sense that the posterior predictive density lets us believe that a wrong model was a correct one. In contrast, the mixed predictive distribution correctly shows the model deficiencies.

We consider grouped data $y^{obs}$ with a covariate generated from the following mixed effects model:

$$y^{obs}_{hj} = x_{hj} \beta + \xi_h + \epsilon_{hj} \quad (h = 1, \ldots, H, j = 1, \ldots, J), \quad (4.6)$$

where $\epsilon_{hj} \sim N(0, \tau^2)$, independent and identically distributed, $\xi_h \sim N(0, \sigma^2)$, independent and identically distributed. I.e., there are $H$ groups of size $J$, each group having a random effect $\xi_h$, and $x_{hj}$ denotes the covariate.

Let us assume that we have fitted a model without the fixed effect $x_{hj}\beta$,

$$y_{hj} = \xi_h + \epsilon_{hj}. \quad (4.7)$$

We now want to use a diagnostic check to investigate whether the model (4.7) is correctly specified, or whether the covariate $x_{hj}$ is needed. This is done by considering the residuals

$$r_{hj} = y^{obs}_{hj} - E[y^{new}_{hj}], \quad (4.8)$$

and by plotting them against $x_{hj}$. In the model (4.7), the incidental variables are $\xi = (\xi_1, \ldots, \xi_H)$ and the primary parameters are $\phi = (\sigma^2, \tau^2)$. In (4.8), we thus obtain two different residuals depending on whether the posterior predictive density or the mixed predictive distribution is used to compute $E[y^{new}]$. In the former case $E[y^{new}_{hj}] = E[\xi_h|y^{obs}]$ and in the latter $E[y^{new}_{hj}] = 0$. 

To be concrete, specify $H = 20$, $J = 10$, $\sigma^2 = 1$, $\tau^2 = 1$, $\beta = 1$, and let the covariate $x = (\bar{x} - \text{mean}(\bar{x}))/\text{sd}(\bar{x})$ where $\bar{x}_{hj} = J(h - 1) + j$. For the variance parameters we use uninformative priors. The model is fitted using a Monte Carlo Markov Chain algorithm (Metropolis et al., 1953; Hastings, 1970).

Figure 4.1 shows the partial residual plots obtained using the posterior predictive density and the mixed predictive distribution. According to the posterior predictive density, there is no discernible misspecification. The mixed predictive distribution, on the other hand, clearly indicates that the covariate is missing in the model. The reason is that the posterior distribution of the random effect adapts well to the data and, in particular, the posterior mean compensates for the missing covariate.

Figure 4.1: Partial residual plots obtained using the posterior predictive density (PPD) and the mixed predictive distribution (MPD).

4.3.2 A spatio-temporal model

Here we consider a spatio-temporal example. See Cressie and Wikle (2011) for an introduction into spatio-temporal statistics. We assume
that we have data generated from:

\[ y_t(s_i) = \xi_t(s_i) + \epsilon_t(s_i), \quad (4.9) \]

where \( \epsilon_t(s_i) \sim N(0, \tau^2) \) is an i.i.d. noise term and where \( \xi_t(s_i) \) is a stationary, isotropic Gaussian process with a separable covariance function

\[ \text{Cov}(\xi_t(s), \xi_t(s')) = \sigma^2 \exp(-|s - s'|/\rho) \exp(-|t - t'|/\theta). \quad (4.10) \]

We then take up the position that we don’t know the true underlying covariance structure and fit a model without any temporal correlation. I.e., we assume the following covariance function for the \( \xi_t(s_i) \) process:

\[ \text{Cov}(\xi_t(s), \xi_t(s')) = \sigma^2 \exp(-|s - s'|/\rho) \delta_{t, t'}. \quad (4.11) \]

Model checking is done by comparing empirical correlations of the observed data with correlations from samples from the posterior predictive distribution or the mixed predictive distribution. As we show below, the posterior predictive distribution does not show model deficiencies in the spatio-temporal dependence structure, which is the critical part of the model. The reason is the following. The goal is to check whether the conditional model for the latent process \( \xi_t(s_i) \), given the primary parameters \( \phi \) but not the observations \( y \), provides a space-time dependence that is in line with the empirical dependence of the data. Now, the posterior distribution of the incidental variables \( \xi_t(s_i) \) will be determined by the observations rather precisely. Therefore, the sets of observations simulated from the posterior predictive distribution will always have a correlation structure that is similar to the correlations in the actual data, even if the correlation model is inadequate.

To be more specific, we simulate from the separable model with the specifications as in (4.10). We use \( n = 20 \) spatial points that are uniformly distributed in \([0, 1]^2\) and \( T = 200 \) equidistant time points. Concerning the primary parameters, we assume \( \rho = 0.6, \sigma^2 = 0.5, \tau^2 = 0.05, \) and \( \theta = -1/\log(0.8) \). Note that the choice of the temporal range parameter \( \theta \) corresponds to an autoregressive coefficient of 0.8. This means that we have strong autocorrelation in the simulated data.
We then fit a model without including any autoregressive term according to (4.11). This model is not appropriate, since we know that the data exhibits temporal autocorrelation. As in the previous example, the model is fitted using an Monte Carlo Markov Chain algorithm. We then simulate from both the posterior predictive density as well as the mixed predictive distribution and compare suitable simulated characteristics with their observed counterparts.

**Figure 4.2:** Comparison of the posterior predictive distribution (PPD) (top row) and the mixed predictive distribution (MPD) (bottom row) for model checking. Simultaneous cross-correlations and lag-1 cross-correlations are plotted against distance. The triangles represent the means of the observed values in each distance class, whereas the dots and the bars are means and 0.025 and 0.975 quantiles of the posterior predictive density and the mixed predictive distribution.

Since our focus lies on modeling spatio-temporal dependencies, we consider simultaneous cross-correlations $\text{corr}(y_t(s_i), y_t(s_j)), i, j = 1, \ldots, n$, and lagged cross-correlations $\text{corr}(y_t(s_i), y_{t-1}(s_j))$. In Figure 4.2, simultaneous and lag-1 cross-correlations versus distance between sites are plotted. The locations are partitioned into distance classes such that in each class there is roughly the same number of stations.

By looking at the posterior predictive distribution plot in Figure 4.2, we observe a very good agreement between the simulated and observed values for both simultaneous and lag-1 correlations. In fact, observed values and simulated expectations are almost indistinguishable. In contrast, the mixed predictive distribution clearly shows that the lag-1 correlations are not modelled appropriately. This is in line with what we desire to happen.
4.4 A case study: space-time rainfall data

In the following, we illustrate the use of the mixed predictive distribution on the basis of a real data set, see Sigrist et al. (2012) for more details. The goal is to generate short term predictions of precipitation using a spatio-temporal model. The question we are trying to answer is whether a relatively simple separable covariance function provides a good fit to the data or whether a more sophisticated space-time dependency structure is needed.

Precipitation in space and time is modeled using a hierarchical Bayesian model with a variant of the Tobit model (Tobin, 1958) at the data stage to account for the fact that the rainfall distribution consists of a discrete point mass at zero (no rainfall) and a continuous distribution above zero (rainfall amount). At the process stage, a spatio-temporal Gaussian process is used to account for structured variation that is not explained by a few covariates. One candidate model is a separable one as in (4.10). On the other hand, Sigrist et al. (2012) propose a physically motivated model that is based on an integro-difference equation which explicitly models physical processes such as advection, diffusion, and convection. Having only a few additional parameters, this extended model allows for non-separability, temporal non-stationarity, and spatial anisotropy in the space-time covariance structure.

As in the previous example, we consider simultaneous cross-correlations 
\[ \text{corr}(y_t(s_i), y_t(s_j)), \ 1 \leq i, j, \leq n \] 
and lagged cross-correlations 
\[ \text{corr}(y_t(s_i), y_{t-k}(s_j)), \ k = 1, 2, 3, \] 
in order to investigate the fit of the two candidate models. Using the posterior predictive distribution (results not shown), we observe that both models provide an apparently excellent fit to the data. The plots based on the mixed predictive distribution are in Figure 4.3. These results show that the separable model is overly simplistic and does not accurately account for the spatio-temporal correlations. The physically motivated model, though, provides a good fit to the spatio-temporal dependence structure. This conclusion is in line with the findings of Sigrist et al. (2012), where both models are used to make short term predictions, and the extended model has a considerably better predictive performance.
Figure 4.3: Comparison of observed and fitted space-time correlations. Simultaneous, lag-1, lag-2, and lag-3 cross-correlations versus distance between sites are plotted. The triangles represent observed values, the dots are means of the mixed predictive distribution (MPD), and the bars show 0.025 and 97.5 quantiles of the mixed predictive distribution. The top plot is based on the separable autoregressive model, and the bottom plot on the physics based model with a more realistic covariance structure.
Bibliography


<table>
<thead>
<tr>
<th>Name</th>
<th>Fabio Roman Albert Sigrist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nationality</td>
<td>Swiss</td>
</tr>
<tr>
<td>City of birth</td>
<td>Zürich</td>
</tr>
<tr>
<td>Date of birth</td>
<td>08.11.1983</td>
</tr>
<tr>
<td>01/2009 - 03/2013</td>
<td>Ph.D. in statistics, ETH Zürich</td>
</tr>
<tr>
<td>09/2011 - 12/2011</td>
<td>Research stay at University of Washington, Seattle</td>
</tr>
<tr>
<td>09/2006 - 02/2008</td>
<td>M.Sc. in mathematics, ETH Zürich</td>
</tr>
<tr>
<td>09/2003 - 08/2006</td>
<td>B.Sc. in mathematics, ETH Zürich / Sapienza - Università di Roma</td>
</tr>
<tr>
<td>07/2003</td>
<td>Matura, Kantonsschule Baden</td>
</tr>
</tbody>
</table>