Journal of the Royal Statistical Society
SERIES B Statistical Methodology
J. R. Statist. Soc. B (2015)

Stochastic partial differential equation based modelling of large space–time data sets

Fabio Sigrist, Hans R. Künsch and Werner A. Stahel

Eidgenössiche Technische Hochschule Zürich, Switzerland

[Received November 2012. Final revision October 2013]

Summary. 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 the solution of a stochastic advection–diffusion partial differential equation provides a flexible model class for spatiotemporal processes which is computationally feasible also for large data sets. The Gaussian process defined through the stochastic partial differential equation has, in general, a non-separable covariance structure. Its parameters can be physically interpreted as explicitly modelling phenomena such as transport and diffusion that occur in many natural processes in diverse fields ranging from environmental sciences to ecology. To obtain computationally efficient statistical algorithms, we use spectral methods to solve the stochastic partial differential equation. 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 cost of Bayesian or frequentist inference being dominated by the fast Fourier transform. The model proposed is applied to post-processing of precipitation forecasts from a numerical weather prediction model for northern Switzerland. In contrast with the raw forecasts from the numerical model, the post-processed forecasts are calibrated and quantify prediction uncertainty. Moreover, they outperform the raw forecasts, in the sense that they have a lower mean absolute error.

Keywords: Advection–diffusion equation; Gaussian process; Numerical weather prediction; Physics-based model; Spatiotemporal model; Spectral methods

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) pointed out, there are two basic paradigms for constructing spatiotemporal 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) referred to this as the 'big n problem'. Factorizing large covariance matrices is not possible without assuming a special structure or using

Address for correspondence: Fabio Sigrist, Seminar für Statistik, Eidgenössiche Technische Hochschule Zürich, Rämistrasse 101, 8092 Zürich, Switzerland.
E-mail: sigrist@stat.math.ethz.ch

 1369–7412/15/77000

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 (Rue and Tjelmeland, 2002; Rue and Held, 2005; Lindgren *et al.*, 2011) or applying tapering (Furrer *et al.*, 2006) thereby obtaining 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) used Fourier functions to reduce computational costs.

In a space–time setting, the situation is the same, if not worse: one runs into a computational bottleneck with high dimensional data since the computational cost to factorize dense $NT \times NT$ covariance matrices is $O\{(NT)^3\}$, N and T being the number of points in space and time respectively. Moreover, specifying flexible and realistic space–time covariance functions is a non-trivial task.

In this paper, we follow the dynamic approach and study models which are defined through a stochastic advection–diffusion partial differential equation (PDE). This has the advantage of providing physically motivated parameterizations of space–time covariances. We show that, when solving the stochastic partial differential equation (SPDE) by using Fourier functions, we 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 those 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, 1962) and Heine (1955). Later works include Jones and Zhang (1997) and Brown *et al.* (2000). Recently, Lindgren *et al.* (2011) have shown how a certain class of SPDEs can be solved by using finite elements to obtain parameterizations of spatial Gaussian Markov random fields. A potential *caveat* of these SPDE approaches is that it is non-trivial to generalize the linear equation to non-linear equations.

Spectral methods for solving PDEs are well established in the numerical mathematics community (see, for example, 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 spatiotemporal statistics, mostly for approximating or solving deterministic integrodifference equations or PDEs. Wikle and Cressie (1999) introduced a dynamic spatiotemporal model obtained from an integrodifference equation that is approximated by using a reduced dimensional spectral basis. Extending this work, Wikle (2002) and Xu *et al.* (2005) proposed parameterizations of spatiotemporal processes based on integrodifference equations. Modelling tropical ocean surface winds, Wikle *et al.* (2001) presented a physics-based model based on the shallow water equations. Cressie and Wikle (2011), chapter 7, gave an overview of basis function expansions in spatiotemporal statistics.

The novel features of our work are the following. Whereas spectral methods have been used for approximating deterministic integrodifference equations and PDEs in the statistical literature, there is no reference, to our knowledge, that explicitly shows how to obtain 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 FFT 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, our main objective is to post-process precipitation forecasts, explicitly modelling spatial and temporal variation. The idea is that the spatiotemporal 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 2 introduces the continuous space–time Gaussian process defined through the advection–diffusion SPDE. In Section 3, it is shown how the solution of the SPDE can be approximated by using the two-dimensional real Fourier transform, and we give convergence rates for the approximation. Next, in Section 4, we show how to do computationally efficient inference. In Section 5, the spatiotemporal model is used as part of a hierarchical Bayesian model, which we then apply to post-processing of precipitation forecasts.

All the methodology that is presented in this paper is implemented in the R package `spate` (see Sigrist *et al.* (2012)).

2. A continuous space–time model: the advection–diffusion stochastic partial differential equation

In one dimension, a fundamental process is the Ornstein–Uhlenbeck process which is governed by a relatively simple stochastic differential equation. The process has an exponential covariance function and its discretized version is the famous auto-regressive AR(1) model. In the two-dimensional spatial case, Whittle (1954) argued convincingly that the process with a Whittle correlation function is an 'elementary' process (see Section 2.2 for further discussion). If the time dimension is added, we think that the process that is defined through the SPDE (1) has properties that make it a good candidate for an elementary spatiotemporal 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 0 (no advection and no diffusion), the covariance structure reduces to a separable covariance structure with an AR(1) structure over time and a certain covariance structure over space.

The advection–diffusion SPDE, which is also called the transport–diffusion SPDE, is given by

$$\frac{\partial}{\partial t}\xi(t,\mathbf{s}) = -\boldsymbol{\mu}^{\mathrm{T}}\nabla\xi(t,\mathbf{s}) + \nabla\cdot\boldsymbol{\Sigma}\nabla\xi(t,\mathbf{s}) - \zeta\,\xi(t,\mathbf{s}) + \varepsilon(t,\mathbf{s}), \tag{1}$$

with $\mathbf{s} = (x, y)^{\mathrm{T}} \in \mathbb{R}^2$, where $\nabla = (\partial/\partial x, \partial/\partial y)^{\mathrm{T}}$ is the gradient operator, and, for a vector field $\mathbf{F} = (F^x, F^y)^{\mathrm{T}}$, $\nabla\cdot\mathbf{F} = \partial F^x/\partial x + \partial F^y/\partial y$ is the divergence operator. $\varepsilon(t,\mathbf{s})$ is a Gaussian process that is temporally white and spatially coloured. See Section 2.2 for a discussion on the choice of the spatial covariance function. Heine (1955) and Whittle (1963) introduced and analysed SPDEs of similar form to equation (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 integrodifference equation models. Without giving any concrete details, Lindgren *et al.* (2011) suggested that this SPDE can be used in connection with their Gaussian Markov random-field method. See also Simpson *et al.* (2012) and Yue *et al.* (2012). Cameletti *et al.* (2013) modelled particulate matter concentration in space and time with a separable covariance structure and an SPDE-based spatial Gaussian Markov random field for the innovation term. Aune and Simpson (2012) and Hu *et al.* (2013) used systems of SPDEs to define multivariate spatial models.

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 $\boldsymbol{\mu}^{\mathrm{T}}\nabla\xi(t,\mathbf{s})$ models transport effects (called advection in weather applications), $\boldsymbol{\mu}=(\mu_x,\mu_y)^{\mathrm{T}}\in\mathbb{R}^2$ being a drift or velocity vector. The second term, $\nabla\cdot\boldsymbol{\Sigma}\nabla\xi(t,\mathbf{s})$, is a diffusion term that can incorporate anisotropy. If $\boldsymbol{\Sigma}$ is the identity matrix, this term reduces to the divergence $\nabla\cdot$ of the gradient ∇ which is the ordinary Laplace operator $\nabla\cdot\nabla=\Delta=\partial^2/\partial x^2+\partial^2/\partial y^2$. The third term $-\zeta\xi(t,\mathbf{s})$, $\zeta>0$, diminishes $\xi(t,\mathbf{s})$ at a constant rate and thus accounts for damping. Finally, $\varepsilon(t,\mathbf{s})$ is a source–sink or stochastic forcing term, which is also called an innovation term, that can be interpreted as describing, among others, convective phenomena in precipitation modelling applications.

Concerning the diffusion matrix $\boldsymbol{\Sigma}$, we suggest the parameterization

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

where $\rho_1>0$, $\gamma>0$ and $\psi\in[0,\pi/2]$. The parameters are interpreted as follows. ρ_1 acts as a range parameter and controls the amount of diffusion. The parameters γ and ψ control the amount and the direction of anisotropy. With $\gamma=1$, isotropic diffusion is obtained.

Fig. 1 illustrates the SPDE (1) and the corresponding PDE without the stochastic innovation term. Figs 1(a)–1(e) show a solution to the PDE which corresponds to the deterministic part of the SPDE that is obtained when there is no stochastic term $\varepsilon(t,\mathbf{s})$. Fig. 1 shows how the initial state in Fig. 1(a) becomes propagated forwards in time. The drift vector points from north-east to south-west and the diffusive part exhibits anisotropy in the same direction. A 100×100 grid is used and the PDE is solved in the spectral domain by using the method that is described below in Section 3. There is a fundamental difference between the deterministic PDE and the probabilistic SPDE. In the first case, a deterministic process is modelled 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. Figs 1(f)–1(j) show one sample from this Gaussian process. The same 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 $\varepsilon(t,\mathbf{s})$, 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.

The use off this spatiotemporal 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 modelled process follows the dynamics of the Ornstein–Uhlenbeck stochastic differential equation. In two dimensions, the same holds true for the process with the Whittle covariance function, and even more so for the process having an exponential covariance structure. Having this in mind, even though the SPDE (1) is physically motivated, it can be used as a general spatiotemporal model. As the case may be, the interpretation of the parameters can be more or less straightforward.

2.1. Spectral density and covariance function

As can be shown by using the Fourier transform (see, for example, Whittle (1963)), if the innovation process $\varepsilon(t,\mathbf{s})$ is stationary with spectral density $\tilde{f}(\mathbf{k})$, the spectrum of the stationary solution $\xi(t,\mathbf{s})$ of the SPDE (1) is

$$f(\omega,\mathbf{k})=\tilde{f}(\mathbf{k})\frac{1}{2\pi}\{(\mathbf{k}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{k}+\zeta)^2+(\omega+\boldsymbol{\mu}^{\mathrm{T}}\mathbf{k})^2\}^{-1}, \tag{3}$$

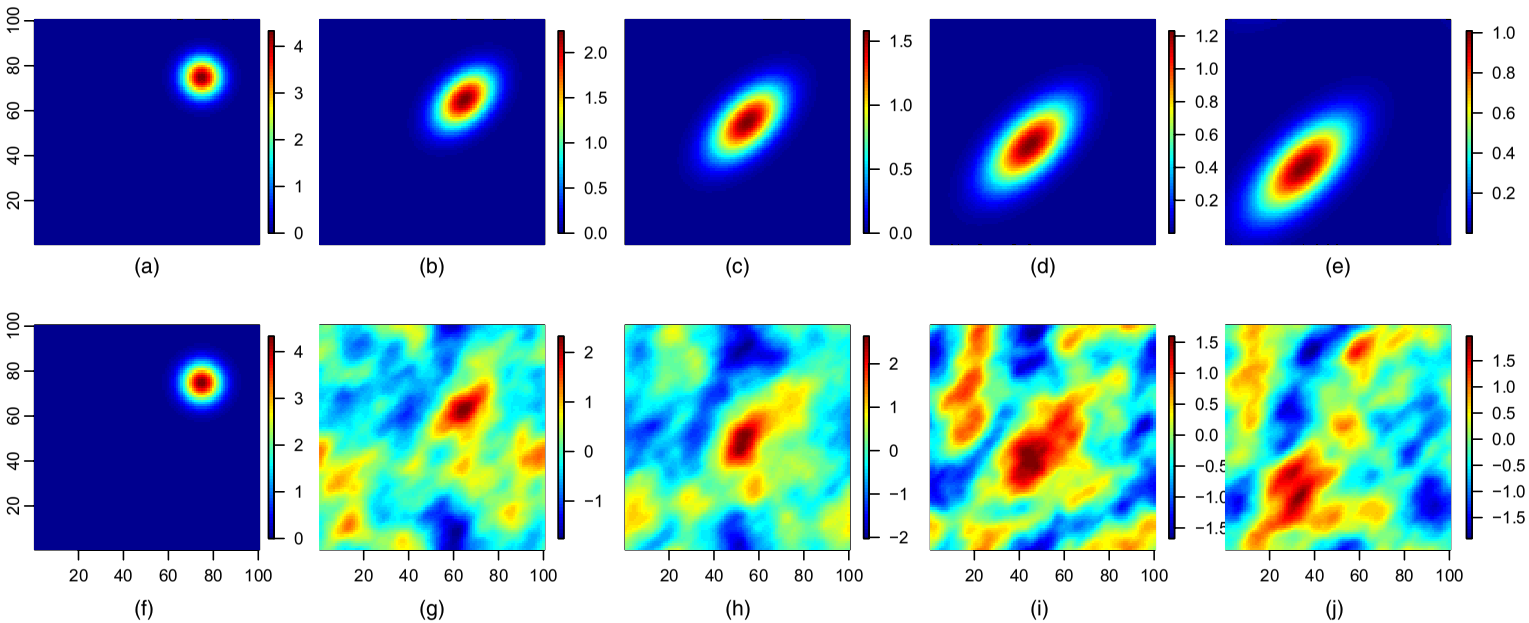

Fig. 1. Illustration of the SPDE (1) and the corresponding PDE (the drift vector points from north-east to south-west and the diffusive part exhibits anisotropy in the same direction; the same parameters are used for both the PDE and the SPDE, i.e. $\zeta = -\log(0.99)$, $\rho_1 = 0.06$, $\gamma = 3$, $\psi = \pi/4$, $\mu_x = -0.1$ and $\mu_y = -0.1$, and for the stochastic innovations, i.e. $\rho_0 = 0.05$ and $\sigma^2 = 0.7^2$; the colour scales are different in different panels): (a)–(e) a solution to the PDE which corresponds to the deterministic part of the SPDE without stochastic term $\varepsilon(t, \mathbf{s})$; (f)–(j) one sample from the distribution specified by the SPDE with a fixed initial condition; (a), (f) $t = 1$; (b), (g) $t = 2$; (c), (h) $t = 3$; (d), (i) $t = 4$; (e), (j) $t = 5$

where $\mathbf{k}$ and ω are spatial wave numbers and temporal frequencies. The covariance function $C(t,\mathbf{s})$ of $\xi(t,\mathbf{s})$ is then given by

$$\begin{aligned} C(t,\mathbf{s}) &= \int f(\omega,\mathbf{k})\exp(\mathrm{i}t\omega)\exp(\mathrm{i}\mathbf{s}'\mathbf{k})\,\mathrm{d}\mathbf{k}\,\mathrm{d}\omega \\ &= \int \tilde{f}(\mathbf{k})\frac{\exp\{-\mathrm{i}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{k}t-(\mathbf{k}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{k}+\zeta)|t|\}}{2(\mathbf{k}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{k}+\zeta)}\exp(\mathrm{i}\mathbf{s}'\mathbf{k})\,\mathrm{d}\mathbf{k}, \end{aligned} \tag{4}$$

where i denotes the imaginary number $\mathrm{i}^2=-1$, and the integration over the temporal frequencies ω 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 $\xi(t,\mathbf{s})$ has a non-separable covariance function (see Gneiting *et al.* (2007b) for a definition of separability). The model reduces to a separable model though, when there is no advection and diffusion, i.e. when both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are 0. In this case, the covariance function is given by $C(t,\mathbf{s})=(1/2\zeta)\exp(-\zeta|t|)C(\mathbf{s})$, where $C(\mathbf{s})$ denotes the spatial covariance function of the innovation process.

2.2. Specification of the innovation process

It is assumed that the innovation process is white in time and spatially coloured. In principle, we can choose any spatial covariance function such that the covariance function in equation (4) is finite at zero. If $\tilde{f}(\mathbf{k})$ is integrable, then $f(\omega,\mathbf{k})$ is also integrable. Similarly to 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) and 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

$$\left(\nabla\cdot\nabla-\frac{1}{\rho_0^2}\right)\varepsilon(t,\mathbf{s})=\mathcal{W}(t,\mathbf{s}), \tag{5}$$

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

$$\tilde{f}(\mathbf{k})=\frac{\sigma^2}{(2\pi)^2}\left(\mathbf{k}^{\mathrm{T}}\mathbf{k}+\frac{1}{\rho_0^2}\right)^{-2}, \qquad \rho_0>0, \quad \sigma>0. \tag{6}$$

The parameter σ^2 determines the marginal variance of $\varepsilon(t,\mathbf{s})$, and ρ_0 is a spatial range parameter.

2.3. Relation to an integrodifference equation

Assuming discrete time steps with lag Δ, Brown *et al.* (2012) considered the integrodifference equation

$$\xi(t,\mathbf{s}) = \exp(-\Delta\zeta) \int_{\mathbb{R}^2} h(\mathbf{s}-\mathbf{s}')\,\xi(t-\Delta,\mathbf{s}')\,\mathrm{d}\mathbf{s}' + \varepsilon(t,\mathbf{s}), \qquad \mathbf{s}\in\mathbb{R}^2, \tag{7}$$

with a Gaussian redistribution kernel

$$h(\mathbf{s}-\mathbf{s}') = (2\pi)^{-1}|2\Delta\boldsymbol{\Sigma}|^{-1/2}\exp\{-(\mathbf{s}-\mathbf{s}'-\Delta\boldsymbol{\mu})^{\mathrm{T}}(2\Delta\boldsymbol{\Sigma})^{-1}(\mathbf{s}-\mathbf{s}'-\Delta\boldsymbol{\mu})/2\},$$

$\varepsilon(t,\mathbf{s})$ being temporally independent and spatially dependent. They showed that, in the limit $\Delta\to 0$, the solution of the integrodifference equation and that of the SPDE (1) coincide. The integrodifference equation is interpreted as follows: the convolution kernel $h(\mathbf{s}-\mathbf{s}')$ determines the weight or the amount of influence that a location $\mathbf{s}'$ at previous time $t-\Delta$ has on the point $\mathbf{s}$ at current time t. This integrodifference equation representation provides an alternative way of interpreting the SPDE model and its parameters. Storvik *et al.* (2002) showed under which conditions a dynamic model determined by an integrodifference equation like equation (7) can be represented by using a parametric joint space–time covariance function, and vice versa. On the basis of the integrodifference equation (7), Sigrist *et al.* (2012) constructed a spatiotemporal model for irregularly spaced data and applied it to obtain short-term predictions of precipitation. Wikle (2002) and Xu *et al.* (2005) also modelled spatiotemporal rainfall on the basis of integrodifference equations.

3. Solution in the spectral space

Solutions $\xi(t,\mathbf{s})$ of the SPDE (1) are defined in continuous space and time. In practice, we need 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 or impossible. However, as we show in what follows, 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) and Cressie and Wikle (2011), chapter 7) approximate the solution $\xi(t,\mathbf{s})$ by a linear combination of deterministic spatial functions $\phi_j(\mathbf{s})$ with random coefficients $\alpha_j(t)$ that evolve dynamically over time:

$$\xi^K(t,\mathbf{s}) = \sum_{j=1}^{K}\alpha_j(t)\,\phi_j(\mathbf{s}) = \boldsymbol{\phi}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\alpha}(t), \tag{8}$$

where $\boldsymbol{\phi}(\mathbf{s}) = (\phi_1(\mathbf{s}),\ldots,\phi_K(\mathbf{s}))^{\mathrm{T}}$ and $\boldsymbol{\alpha}(t) = (\alpha_1(t),\ldots,\alpha_K(t))^{\mathrm{T}}$. To be more specific, we use Fourier functions

$$\phi_j(\mathbf{s}) = \exp(\mathrm{i}\mathbf{k}_j^{\mathrm{T}}\mathbf{s}), \tag{9}$$

where $\mathbf{k}_j = (k_j^x, k_j^y)^{\mathrm{T}}$ is a spatial wave number.

The advantages of using Fourier functions for solving linear deterministic PDEs are well known; see, for example, 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 when all the basis functions are retained. 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 SPDE (1): if the initial condition and the innovation process are in the space that is spanned by a finite number of

Fourier functions, then the solution of the SPDE (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

$$\begin{aligned} \xi^K(0,\mathbf{s}) &= \boldsymbol{\phi}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\alpha}(0), \\ \varepsilon^K(t,\mathbf{s}) &= \boldsymbol{\phi}(\mathbf{s})^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}}(t) \end{aligned} \tag{10}$$

where $\boldsymbol{\phi}(\mathbf{s}) = (\phi_1(\mathbf{s}), \ldots, \phi_K(\mathbf{s}))^{\mathrm{T}}$, $\phi_j(\mathbf{s})$ is given in equation (9), $\boldsymbol{\alpha}(0) \sim N[\mathbf{0}, \mathrm{diag}\{\tilde{f}_0(\mathbf{k}_j)\}]$, $\tilde{f}_0(\cdot)$ being a spectral density, and $\tilde{\boldsymbol{\varepsilon}}(t)$ is K-dimensional Gaussian white noise independent of $\boldsymbol{\alpha}(0)$ with

$$\mathrm{cov}\{\tilde{\boldsymbol{\varepsilon}}(t), \tilde{\boldsymbol{\varepsilon}}(t')\} = \delta_{t,t'}\,\mathrm{diag}\{\tilde{f}(\mathbf{k}_j)\}, \tag{11}$$

where $\tilde{f}(\cdot)$ is a spectral density and $\delta_{t,t'}$ the Kronecker delta function equalling 1 if $t = t'$ and 0 otherwise. Then the process $\xi^K(t,\mathbf{s}) = \boldsymbol{\phi}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\alpha}(t)$, where the components $\alpha_j(t)$ are given by

$$\alpha_j(t) = \exp(h_j t)\,\alpha_j(0) + \int_0^t \exp\{h_j(t-u)\}\,\tilde{\varepsilon}_j(u)\,\mathrm{d}u, \tag{12}$$

with $h_j = -\mathrm{i}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{k}_j - \mathbf{k}_j^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{k}_j - \zeta$, is a solution of the SPDE (1). For $t \to \infty$, the influence of the initial condition $\exp(h_j t)\,\alpha_j(0)$ converges to zero and the process $\xi^K(t,\mathbf{s})$ converges to a time stationary Gaussian process with mean 0 and

$$\mathrm{cov}\{\xi^K(t+\Delta t,\mathbf{s}), \xi^K(t,\mathbf{s}')\} = \boldsymbol{\phi}(\mathbf{s})^{\mathrm{T}}\mathrm{diag}\left\{\frac{-\exp(h_j\Delta t)\,\tilde{f}(\mathbf{k}_j)}{h_j + h_j^*}\right\}\boldsymbol{\phi}(\mathbf{s}')^*,$$

where the asterisk denotes complex conjugation.

This result shows that the solution of the SPDE is exact over time, given the frequencies included. In contrast with finite differences, one does not accumulate errors over time. 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. The approximation error of $\xi^K(t,\mathbf{s})$ to the space–time stationary solution of the SPDE (1) depends on only the number of spectral terms and not on the temporal discretization; see also proposition 2 below. Since Fourier terms are global functions, stationarity in space, but not in time, is a necessary assumption.

Proof. By equation (12), we have

$$\frac{\partial}{\partial t}\xi^K(t,\mathbf{s}) = \sum_{j=1}^K \dot{\alpha}_j(t)\,\phi_j(\mathbf{s}) = \sum_{j=1}^K \{h_j\alpha_j(t) + \tilde{\varepsilon}_j(t)\}\,\phi_j(\mathbf{s}).$$

In contrast, since the functions $\phi_j(\mathbf{s}) = \exp(\mathrm{i}\mathbf{k}_j^{\mathrm{T}}\mathbf{s})$ are Fourier terms, differentiation in the physical space corresponds to multiplication in the spectral space:

$$\boldsymbol{\mu}^{\mathrm{T}}\nabla\phi_j(\mathbf{s}) = \mathrm{i}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{k}_j\,\phi_j(\mathbf{s}) \tag{13}$$

and

$$\nabla\cdot\boldsymbol{\Sigma}\nabla\phi_j(\mathbf{s}) = -\mathbf{k}_j^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{k}_j\,\phi_j(\mathbf{s}). \tag{14}$$

Therefore, by the definition of h_j,

$$(-\boldsymbol{\mu}^{\mathrm{T}}\nabla + \nabla\cdot\boldsymbol{\Sigma}\nabla - \zeta)\sum_{j=1}^K \alpha_j(t)\,\phi_j(\mathbf{s}) = \sum_{j=1}^K h_j\alpha_j(t)\,\phi_j(\mathbf{s}).$$

Together, we have

$$\frac{\partial}{\partial t}\xi^K(t,\mathbf{s}) = (-\boldsymbol{\mu}^{\mathrm{T}}\nabla + \nabla\cdot\boldsymbol{\Sigma}\nabla - \zeta)\,\xi^K(t,\mathbf{s}) + \varepsilon^K(t,\mathbf{s})$$

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,

$$\begin{aligned}\lim_{t\to\infty} \operatorname{cov}\{\alpha_j(t+\Delta t), \alpha_{j'}(t)\} &= \lim_{t\to\infty} \exp(h_j \Delta t)\delta_{j,j'}\,\tilde{f}(\mathbf{k}_j)\int_0^t \exp\{-(h_j + h_{j'}^*)(t-u)\}\,\mathrm{d}u \\ &= -\frac{\exp(h_j\Delta t)}{h_j + h_j^*}\delta_{j,j'}\,\tilde{f}(\mathbf{k}_j), \end{aligned} \tag{15}$$

and thus the last statement follows.

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

$$\varepsilon(t,\mathbf{s}) = \int \exp(\mathrm{i}\mathbf{k}^{\mathrm{T}}\mathbf{s})\,\mathrm{d}\tilde{\varepsilon}_t(\mathbf{k})$$

where $\tilde{\varepsilon}_t$ has orthogonal increments $\operatorname{cov}\{\mathrm{d}\tilde{\varepsilon}_t(\mathbf{k}), \mathrm{d}\tilde{\varepsilon}_{t'}(\mathbf{l})\} = \delta_{t,t'}\delta_{\mathbf{k},\mathbf{l}}\tilde{f}(\mathbf{k})$ and $\tilde{f}(\cdot)$ is the spectral density of $\varepsilon(t,\cdot)$ (see, for example, Cramér and Leadbetter (1967)). This implies that we can approximate any stationary field, in particular also the field with a Whittle covariance function, by a finite linear combination of complex exponentials, and the covariance of $\tilde{\boldsymbol{\varepsilon}}(t)$ is a diagonal matrix as required in proposition 1. Its entries are specified in expression (6). Concerning the initial state, we can use the stationary distribution of $\xi(t,\cdot)$. An alternative choice is to use the same spatial distribution as for the innovations: $\tilde{f}_0(\cdot) = \tilde{f}(\cdot)$.

3.1. Approximation bound

By passing to the limit $K \to \infty$ such that both the wave numbers $\mathbf{k}_j$ cover the entire domain $\mathbb{R}^2$ and the distance between neighbouring wave numbers goes to 0, we obtain from equation (8) the stationary (in space and time) solution with spectral density as in equation (3). In practice, if we use the discrete Fourier transform, or its fast variant, the FFT, the wave numbers 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.

On the basis of these considerations, we assume, in what follows, that $\mathbf{s} \in [0,1]^2$ 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 $\mathbf{s} \in [0,0.5]^2$ and then embed it in $[0,1]^2$. As in the discrete Fourier transform, if we choose $\mathbf{s} \in [0,1]^2$, it follows that the spatial wave numbers $\mathbf{k}_j$ lie on the $n \times n$ grid given by $D_n = \{2\pi(i,j) : -(n/2+1) \leqslant i,j \leqslant n/2\} = \{-2\pi(n/2+1), \ldots, 2\pi n/2\}^2$ with $n^2 = K$, n being an even natural number. We then have the following convergence result.

Proposition 2. When $K \to \infty$, the approximation $\xi^K(t,\mathbf{s})$ converges in law to the solution $\xi(t,\mathbf{s})$ of the SPDE (1) with $\mathbf{s} \in [0,1]^2$ wrapped on a torus, and we have the bound

$$|C(t,\mathbf{s}) - C^K(t,\mathbf{s})| \leqslant \sigma_\xi^2 - \sigma_{\xi^K}^2, \tag{16}$$

where $C(t,\mathbf{s})$ and $C^K(t,\mathbf{s})$ denote the covariance functions of $\xi(t,\mathbf{s})$ and $\xi^K s(t,\mathbf{s})$ respectively,

and where $\sigma_\xi^2 = C(0, \mathbf{0})$ and $\sigma_{\xi^K}^2 = C^K(0, \mathbf{0})$ denote the marginal variances of these two processes.

Proof. Similarly to expression (4) and because $\mathbf{k} \in 2\pi\mathbb{Z}^2$, it follows that the covariance function of $\xi(t, \mathbf{s})$ is given by

$$\begin{aligned} C(t, \mathbf{s}) &= \sum_{\mathbf{k} \in 2\pi\mathbb{Z}^2} \int f(\omega, \mathbf{k}) \exp(\mathrm{i}t\omega)\, \mathrm{d}\omega \exp(\mathrm{i}\mathbf{s}'\mathbf{k}) \\ &= \sum_{\mathbf{k} \in 2\pi\mathbb{Z}^2} \tilde{f}(\mathbf{k}) \frac{-\exp(h_\mathbf{k} t)}{h_\mathbf{k} + h_\mathbf{k}^*} \exp(\mathrm{i}\mathbf{s}'\mathbf{k}), \end{aligned} \tag{17}$$

where $h_\mathbf{k} = -\mathrm{i}\boldsymbol{\mu}^\mathrm{T}\mathbf{k} - \mathbf{k}^\mathrm{T}\boldsymbol{\Sigma}\mathbf{k} - \zeta$. From proposition 1 we know that the approximate solution $\xi^K(t, \mathbf{s})$ has the covariance function

$$C^K(t, \mathbf{s}) = \sum_{\mathbf{k} \in D_n} \tilde{f}(\mathbf{k}) \frac{-\exp(h_\mathbf{k} t)}{h_\mathbf{k} + h_\mathbf{k}^*} \exp(\mathrm{i}\mathbf{s}'\mathbf{k}). \tag{18}$$

It follows that

$$\begin{aligned} |C(t, \mathbf{s}) - C^K(t, \mathbf{s})| &= \left| \sum_{\mathbf{k} \in 2\pi\mathbb{Z}^2} \tilde{f}(\mathbf{k}) \frac{-\exp(h_\mathbf{k} t)}{h_\mathbf{k} + h_\mathbf{k}^*} (1 - \mathbb{1}_{\{\mathbf{k} \in D_n\}}) \exp(\mathrm{i}\mathbf{s}'\mathbf{k}) \right| \\ &\leqslant \sum_{\mathbf{k} \in 2\pi\mathbb{Z}^2} \tilde{f}(\mathbf{k}) \frac{-1}{h_\mathbf{k} + h_\mathbf{k}^*} (1 - \mathbb{1}_{\{\mathbf{k} \in D_n\}}) \\ &= \sigma_\xi^2 - \sigma_{\xi^K}^2. \end{aligned} \tag{19}$$

Not surprisingly, this result tells us that the rate of convergence essentially depends on the smoothness properties of the process $\xi(t, \mathbf{s})$, i.e. on how fast the spectrum decays. The smoother $\xi(t, \mathbf{s})$, i.e. 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 (1) with $\mathbf{s} \in \mathbb{R}^2$ and the periodic solution with $\mathbf{s} \in [0, 1]^2$ wrapped on a torus. For notational simplicity, we have denoted both of them by $\xi(t, \mathbf{s})$. The finite dimensional solution $\xi^K(t, \mathbf{s})$ is an approximation to both of the above infinite dimensional solutions. The above convergence result, though, holds true only for the solution on the torus.

3.2. Real Fourier functions and discretization in time and space

To apply the model to real data, we must discretize it. In what follows, we consider the process $\xi(t, \mathbf{s})$ on a regular grid of $n \times n = N$ spatial locations $\mathbf{s}_1, \ldots, \mathbf{s}_N$ in $[0, 1]^2$ and at equidistant time points $t_1, \ldots, t_T$ with $t_i - t_{i-1} = \Delta$. These two assumptions can be easily relaxed, i.e. we 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 4.2). The latter can be done by taking a time varying Δ.

For illustration, we have stated the results in the previous section by using complex Fourier functions. However, when discretizing the model, we obtain a linear Gaussian state space model with a propagator matrix $\mathbf{G}$ that contains complex numbers, owing to equation (13). To avoid this, we replace the complex terms $\exp(\mathrm{i}\mathbf{k}_j^\mathrm{T}\mathbf{s})$ with real $\cos(\mathbf{k}_j^\mathrm{T}\mathbf{s})$ and $\sin(\mathbf{k}_j^\mathrm{T}\mathbf{s})$ 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

Fourier transform. For notational simplicity, we shall drop the superscript 'K' from $\xi^K(t,\mathbf{s})$. 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, with initial state $\boldsymbol{\alpha}(t_0) \sim N(0, \tilde{\mathbf{Q}}_0)$, $\tilde{\mathbf{Q}}_0$ diagonal, a stationary solution of the SPDE (1) is of the form

$$\boldsymbol{\xi}(t_{i+1}) = \boldsymbol{\Phi}\,\boldsymbol{\alpha}(t_{i+1}), \tag{20}$$

$$\boldsymbol{\alpha}(t_{i+1}) = \mathbf{G}\,\boldsymbol{\alpha}(t_i) + \tilde{\boldsymbol{\varepsilon}}(t_{i+1}), \qquad \tilde{\boldsymbol{\varepsilon}}(t_{i+1}) \sim N(0, \tilde{\mathbf{Q}}), \tag{21}$$

with stacked vectors $\boldsymbol{\xi}(t_i) = (\xi(t_i,\mathbf{s}_1), \ldots, \xi(t_i,\mathbf{s}_N))^{\mathrm{T}}$ and cosine and sine coefficients $\boldsymbol{\alpha}(t_i) = (\alpha_1^{(\mathrm{c})}(t_i), \ldots, \alpha_4^{(\mathrm{c})}(t_i), \alpha_5^{(\mathrm{c})}(t_i), \alpha_5^{(\mathrm{s})}(t_i), \ldots, \alpha_{K/2+2}^{(\mathrm{c})}(t_i), \alpha_{K/2+2}^{(\mathrm{s})}(t_i))^{\mathrm{T}}$, where $\boldsymbol{\Phi}$ applies the discrete, real Fourier transformation, $\mathbf{G}$ is a block diagonal matrix with 2×2 blocks and $\tilde{\mathbf{Q}}$ is a diagonal matrix. These matrices are defined as follows.

(a)
$$\boldsymbol{\Phi} = (\boldsymbol{\phi}(\mathbf{s}_1), \ldots, \boldsymbol{\phi}(\mathbf{s}_N))^{\mathrm{T}},$$
$$\boldsymbol{\phi}(\mathbf{s}_l) = (\phi_1^{(\mathrm{c})}(\mathbf{s}_l), \ldots, \phi_4^{(\mathrm{c})}(\mathbf{s}_l), \phi_5^{(\mathrm{c})}(\mathbf{s}_l), \phi_5^{(\mathrm{s})}(\mathbf{s}_l), \ldots, \phi_{K/2+2}^{(\mathrm{c})}(\mathbf{s}_l), \phi_{K/2+2}^{(\mathrm{s})}(\mathbf{s}_l))^{\mathrm{T}},$$
$$\phi_j^{(\mathrm{c})}(\mathbf{s}_l) = \cos(\mathbf{k}_j^{\mathrm{T}}\mathbf{s}_l), \qquad \phi_j^{(\mathrm{s})}(\mathbf{s}_l) = \sin(\mathbf{k}_j^{\mathrm{T}}\mathbf{s}_l), \qquad l = 1, \ldots, n^2;$$

(b)
$$(\mathbf{G})_{1:4,1:4} = \mathrm{diag}[\exp\{-\Delta(\mathbf{k}_j^{\mathrm{T}}\boldsymbol{\Sigma}k_j + \zeta)\}],$$
$$(\mathbf{G})_{5:K,5:K} = \mathrm{diag}[\exp\{-\Delta(\mathbf{k}_j^{\mathrm{T}}\boldsymbol{\Sigma}k_j + \zeta)\}\{\cos(\Delta\boldsymbol{\mu}^{\mathrm{T}}\mathbf{k}_j)\mathbf{1}_2 - \sin(\Delta\boldsymbol{\mu}^{\mathrm{T}}\mathbf{k}_j)\mathbf{J}_2\}],$$
where
$$\mathbf{1}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{J}_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{22}$$

(c)
$$\tilde{\mathbf{Q}} = \mathrm{diag}(\tilde{f}(\mathbf{k}_j)[1 - \exp\{-2\Delta(\mathbf{k}_j^{\mathrm{T}}\boldsymbol{\Sigma}k_j + \zeta)\}/\{2(\mathbf{k}_j^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{k}_j + \zeta)\}]),$$

(d)
$$\tilde{\mathbf{Q}}_0 = (\mathbf{1}_N - \mathbf{G}\mathbf{G}^{\mathrm{T}})^{-1}\tilde{\mathbf{Q}}.$$

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

$$\begin{aligned}\xi(t,\mathbf{s}_l) &= \sum_{j=1}^{4} \alpha_j^{(\mathrm{c})}(t)\,\phi_j^{(\mathrm{c})}(\mathbf{s}_l) + \sum_{j=5}^{K/2+2} \{\alpha_j^{(\mathrm{c})}(t)\,\phi_j^{(\mathrm{c})}(\mathbf{s}_l) + \alpha_j^{(\mathrm{s})}(t)\,\phi_j^{(\mathrm{s})}(\mathbf{s}_l)\} \\ &= \boldsymbol{\phi}(\mathbf{s}_l)^{\mathrm{T}}\,\boldsymbol{\alpha}(t),\end{aligned} \tag{23}$$

and the Fourier coefficients $\boldsymbol{\alpha}(t)$ evolve dynamically over time according to the vector auto-regression (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 wave numbers $\mathbf{k}_j$, the sine terms equal 0 on the grid, i.e. $\sin(\mathbf{k}_j^{\mathrm{T}}\mathbf{s}_l) = 0$, for all $l = 1, \ldots, n^2$ and $\mathbf{k}_j \in \{(0,0)^{\mathrm{T}}, (0,n\pi)^{\mathrm{T}}, (n\pi,0)^{\mathrm{T}}, (n\pi,n\pi)^{\mathrm{T}}\}$ (Fig. 2). Equations (20) and (21) form a linear Gaussian state space model with parametric propagator matrix $\mathbf{G}$ and innovation covariance matrix $\tilde{\mathbf{Q}}$, the parameterization being determined by the corresponding SPDE.

Model (20)–(21) is similar to that discussed in Cressie and Wikle (2011), chapter 7, but the derivation as an exact solution to the SPDE (1) rather than a deterministic PDE is different.

Proof. Similarly to proposition 1, we first derive the continuous time solution. Using

$$\boldsymbol{\mu}^{\mathrm{T}}\nabla\phi_j^{(\mathrm{c})}(\mathbf{s}_l) = -\boldsymbol{\mu}^{\mathrm{T}}\mathbf{k}_j\,\phi_j^{(\mathrm{s})}(\mathbf{s}_l),$$
$$\boldsymbol{\mu}^{\mathrm{T}}\nabla\phi_j^{(\mathrm{s})}(\mathbf{s}_l) = \boldsymbol{\mu}^{\mathrm{T}}\mathbf{k}_j\,\phi_j^{(\mathrm{c})}(\mathbf{s}_l),$$

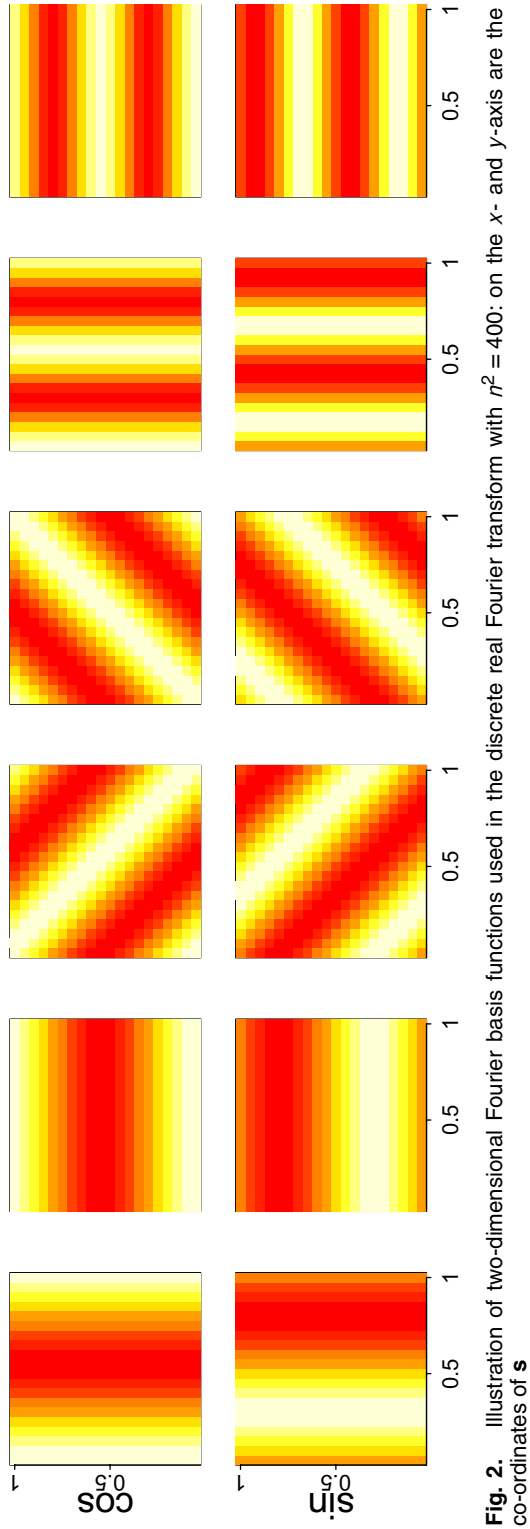

Fig. 2. Illustration of two-dimensional Fourier basis functions used in the discrete real Fourier transform with $n^2 = 400$: on the x- and y-axis are the co-ordinates of $\mathbf{s}$

$$\nabla \cdot \boldsymbol{\Sigma} \nabla \phi_j^{(\mathrm{c})}(\mathbf{s}_l) = -\mathbf{k}_j^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{k}_j \phi_j^{(\mathrm{c})}(\mathbf{s}_l),$$

$$\nabla \cdot \boldsymbol{\Sigma} \nabla \phi_j^{(\mathrm{s})}(\mathbf{s}_l) = -\mathbf{k}_j^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{k}_j \phi_s^{(\mathrm{s})} j(\mathbf{s}_l),$$

and the same arguments as in the proof of proposition 1, it follows that the continuous time solution is of the form (23). For each pair of cosine–sine coefficients $\boldsymbol{\alpha}_j(t) = (\alpha_j^{(\mathrm{c})}(t), \alpha_j^{(\mathrm{s})}(t))^{\mathrm{T}}$ we have

$$\boldsymbol{\alpha}_j(t) = \exp(\mathbf{H}_j t)\, \boldsymbol{\alpha}_j(0) + \int_0^t \exp\{\mathbf{H}_j(t-u)\}\, \tilde{\boldsymbol{\varepsilon}}_j(u)\, \mathrm{d}u, \tag{24}$$

where

$$\mathbf{H}_j = \begin{pmatrix} -\mathbf{k}_j^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{k}_j - \zeta & -\boldsymbol{\mu}^{\mathrm{T}} \mathbf{k}_j \\ \boldsymbol{\mu}^{\mathrm{T}} \mathbf{k}_j & -\mathbf{k}_j^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{k}_j - \zeta \end{pmatrix}.$$

Now $\mathbf{H}_j$ can be written as

$$\mathbf{H}_j = (-\mathbf{k}_j^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{k}_j - \zeta)\mathbf{1}_2 - \boldsymbol{\mu}^{\mathrm{T}} \mathbf{k}_j \mathbf{J}_2,$$

where

$$\mathbf{1}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\mathbf{J}_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Since $\mathbf{1}_2$ and $\mathbf{J}_2$ commute, we have

$$\begin{aligned} \exp(\mathbf{H}_j t) &= \exp\{-t(\mathbf{k}_j^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{k}_j + \zeta)\mathbf{1}_2\} \exp(-t\boldsymbol{\mu}^{\mathrm{T}} \mathbf{k}_j J_2) \\ &= \exp\{-t(\mathbf{k}_j^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{k}_j + \zeta)\}\{\cos(t\boldsymbol{\mu}^{\mathrm{T}} \mathbf{k}_j)\mathbf{1}_2 - \sin(t\boldsymbol{\mu}^{\mathrm{T}} \mathbf{k}_j)\mathbf{J}_2\}. \end{aligned} \tag{25}$$

For the calculation of the exponential function of the matrix $\mathbf{J}_2$, see, for example Bronson and Costa (2007), chapter 4.

Analogously, we derive for the first four cosine terms

$$\alpha_j^{\mathrm{c}}(t) = \exp\{-(\mathbf{k}_j^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{k}_j + \zeta)t\}\, \alpha_j^{\mathrm{c}}(0) + \int_0^t \exp\{-(\mathbf{k}_j^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{k}_j + \zeta)(t-u)\}\, \tilde{\varepsilon}_j(u)\, \mathrm{d}u, \qquad j = 1, \ldots, 4. \tag{26}$$

Expressions (25) and (26) give the propagator matrix $\mathbf{G}$.

For the discrete time solution, in addition to the propagation

$$\boldsymbol{\alpha}_j(t+\Delta) = \exp(\mathbf{H}_j \Delta)\, \boldsymbol{\alpha}_j(t),$$

we need to calculate the covariance of the integrated stochastic innovation term

$$\int_t^{t+\Delta} \exp\{\mathbf{H}_j(t+\Delta-u)\}\, \tilde{\boldsymbol{\varepsilon}}_j(u)\, \mathrm{d}u.$$

This is calculated as

$$\begin{aligned} &\int_t^{t+\Delta} \exp\{\mathbf{H}_j(t+\Delta-u)\}\, \tilde{f}(\mathbf{k}_j) \exp\{\mathbf{H}_j'(t+\Delta-u)\}\, \mathrm{d}u \\ &\qquad = \int_0^{\Delta} \exp\{\mathbf{H}_j(\Delta-u)\}\, \tilde{f}(\mathbf{k}_j) \exp\{\mathbf{H}_j'(\Delta-u)\}\, \mathrm{d}u \\ &\qquad = \int_0^{\Delta} \tilde{f}(\mathbf{k}_j) \exp\{-2(\mathbf{k}_j^{\mathrm{T}} \boldsymbol{\Sigma} k_j + \zeta)(\Delta-u)\}\mathbf{1}_2\, \mathrm{d}u \end{aligned}$$

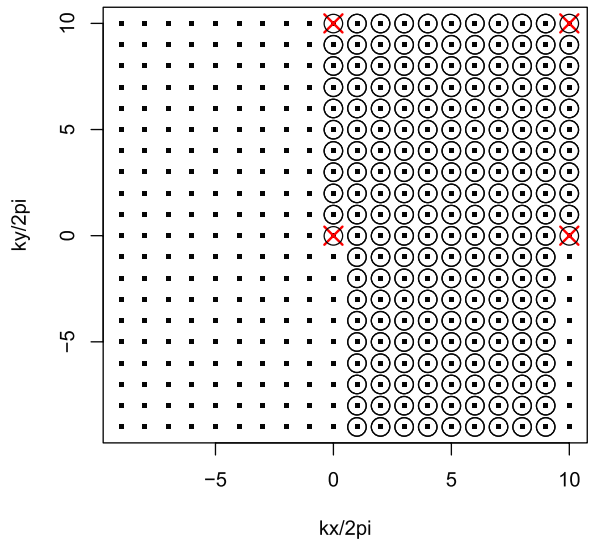

Fig. 3. Illustration of spatial wave numbers for the two-dimensional discrete real Fourier transform with $n^2 = 400$ grid points

$$= \tilde{f}(\mathbf{k}_j) \frac{1 - \exp\{-2(\mathbf{k}_j^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{k}_j + \zeta)\Delta\}}{2(\mathbf{k}_j^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{k}_j + \zeta)} \mathbf{1}_2.$$

For the first four cosine terms, calculations are done analogously. The covariance matrix $\tilde{\mathbf{Q}}_0$ of the initial state $\boldsymbol{\alpha}(t_0)$ is assumed to be the covariance matrix of the stationary distribution of $\boldsymbol{\alpha}(t_i)$. $\tilde{\mathbf{Q}}_0$ is diagonal since $\mathbf{G}\mathbf{G}^{\mathrm{T}}$ is diagonal; see the proof of algorithm 1 in Section 4.1. This then gives result (20)–(21).

The discrete complex Fourier transform uses n^2 different wave numbers $\mathbf{k}_j$, each having a corresponding Fourier term $\exp(\mathrm{i}\mathbf{k}_j^{\mathrm{T}}\mathbf{s})$. The real Fourier transform, in contrast, uses $n^2/2+2$ different wave numbers, 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). Fig. 3 illustrates an example of the spatial wave numbers, with $n^2 = 20 \times 20 = 400$ grid points. The dots with a circle represent the wave numbers that are actually used in the real Fourier transform, and the red crosses mark the wave numbers having only a cosine term. Note that in equation (23) we choose to order the spatial wave numbers such that the first four spatial wave numbers correspond to the cosine-only terms. To obtain an idea of what the basis functions $\cos(\mathbf{k}_j^{\mathrm{T}}\mathbf{s})$ and $\sin(\mathbf{k}_j^{\mathrm{T}}\mathbf{s})$ look like, we plot in Fig. 2 12 low frequency basis functions corresponding to the six spatial frequencies that are closest to the origin $\mathbf{0}$. Further, in Fig. 4, there is an example of a propagator matrix $\mathbf{G}$ when $n = 4$, i.e. when 16 (4^2) spatial basis functions are used. The upper left-hand 4×4 diagonal matrix corresponds to the cosine-only frequencies. The 2×2 blocks following correspond to wave numbers 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 $\boldsymbol{\alpha}(t)$ at each time t. Furthermore, N denotes the number of points at which the process $\boldsymbol{\xi}(t)$ is modelled, and n is the number of points on each axis of

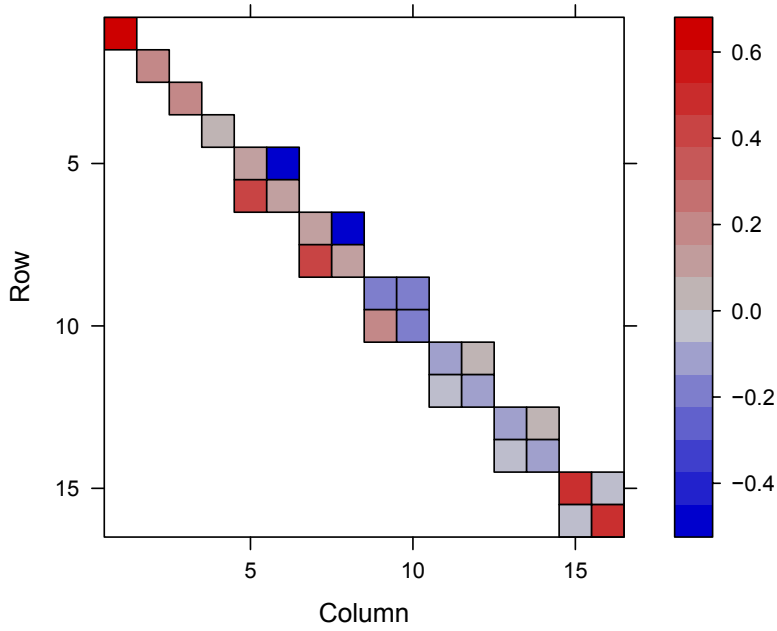

Fig. 4. Illustration of propagator matrix **G**: 16 real Fourier functions are used ($n = 4$)

the quadratic grid used. Often, we have $n^2 = N = K$. However, if we use a reduced dimensional Fourier basis, K is smaller than N; see Section 4.2.

3.3. Remarks on finite differences

Another approach to solve PDEs or SPDEs such as equation (1) consists of using a discretization such as finite differences. Stroud *et al.* (2010) used 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) stated, 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, numerical stability conditions need to be fulfilled so that the approximate solution is meaningful. Since these conditions depend on the values of the unknown parameters, we can run into problems.

In addition, computational tractability is an issue. In fact, we have tried to solve the SPDE (1) by using finite differences as described in what follows. A finite difference approximation in equation (1) leads to a vector auto-regressive model with a sparse propagator matrix being determined by the discretization. The innovation term ε can be approximated by 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, for example, 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.

4. Computationally efficient statistical inference

The computational cost for one evaluation of the likelihood or one sample from the full conditional in a spatiotemporal model with T time points and N spatial points equals $O\{(NT)^3\}$ when taking a naive approach. Using the Kalman filter or the FFBS algorithm (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994), depending on what is needed, this cost is reduced to $O(TN^3)$ which, generally, is still too high for large data sets. In what follows, we show how evaluation of the likelihood and sampling from the full conditional of the latent process 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 FFT (Cooley and Tukey, 1965) which are $O\{TN\log(N)\}$. Furthermore, the 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 $\nu(t_{i+1},\mathbf{s}) \sim^{\text{IID}} N(0,\tau^2)$ to expression (20) to account for small-scale variation and/or measurement errors. In geostatistics, this term is called the nugget effect. Denoting the observations at time t_i by $\mathbf{w}(t_i)$, we then have the linear Gaussian state space model

$$\begin{aligned} \mathbf{w}(t_{i+1}) &= \boldsymbol{\Phi}\,\boldsymbol{\alpha}(t_{i+1}) + \boldsymbol{\nu}(t_{i+1}), & \boldsymbol{\nu}(t_{i+1}) &\sim N(0,\tau^2\mathbf{1}_N), \\ \boldsymbol{\alpha}(t_{i+1}) &= \mathbf{G}\,\boldsymbol{\alpha}(t_i) + \tilde{\boldsymbol{\varepsilon}}(t_{i+1}), & \tilde{\boldsymbol{\varepsilon}}(t_{i+1}) &\sim N(0,\tilde{\mathbf{Q}}). \end{aligned} \tag{27}$$

Note that $\boldsymbol{\xi}(t_{i+1}) = \boldsymbol{\Phi}\,\boldsymbol{\alpha}(t_{i+1})$. As mentioned before, irregular spatial data can be modelled by adopting a data augmentation approach (see Sigrist *et al.* (2012)) or by using an incidence matrix (see Section 4.2). For 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 (Wikle *et al.*, 1998) and specifying a non-Gaussian distribution for $\mathbf{w}|\boldsymbol{\xi}$. The posterior can then no longer be evaluated exactly. But approximate posterior probabilities can still be computed by using, for instance, simulation-based methods such as Markov chain Monte Carlo (MCMC) sampling (see, for example, Gilks *et al.* (1996) or Robert and Casella (2004)). An additional advantage of Bayesian hierarchical models is that these models can be extended, for instance, to account for temporal non-stationarity by letting one or several parameters vary over time.

4.1. Kalman filtering and backward sampling in the spectral space

When following both a frequentist or a Bayesian paradigm, it is crucial that one can evaluate the likelihood of the hyperparameters given $\mathbf{w}$ with a reasonable computational effort. In addition, when doing Bayesian inference, one needs to be able to simulate efficiently from the full conditional of the latent process $[\boldsymbol{\xi}|\cdot]$, or, equivalently, the Fourier coefficients $[\boldsymbol{\alpha}|\cdot]$. 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, we 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 $\boldsymbol{\alpha}$.

Algorithm 1 in Table 1 shows the Kalman filter in the spectral space. For simplicity, we assume

Table 1. Algorithm 1: spectral Kalman filter

Input: T, $\tilde{\mathbf{w}}$, $\mathbf{G}$, τ^2, $\tilde{\mathbf{Q}}$, $\mathbf{F}$
Output: forecast and filter means $\mathbf{m}_{t_i|t_{i-1}}$ and $\mathbf{m}_{t_i|t_i}$, and covariance matrices $\mathbf{R}_{t_i|t_i}$ and $\mathbf{R}_{t_i|t_{i-1}}$, $i = 1, \ldots, T$
$\mathbf{m}_{t_0|t_0} = \mathbf{0}$
$\mathbf{R}_{t_0|t_0} = \tilde{\mathbf{Q}}$
for $i = 1, \ldots, T$ do
$\mathbf{m}_{t_i|t_{i-1}} = \mathbf{G}\mathbf{m}_{t_{i-1}|t_{i-1}}$
$\mathbf{R}_{t_i|t_{i-1}} = \tilde{\mathbf{Q}} + \mathbf{R}_{t_{i-1}|t_{i-1}}\mathbf{F}$
$\mathbf{R}_{t_i|t_i} = (\tau^{-2}\mathbf{1}_N + \mathbf{R}^{-1}_{t_i|t_{i-1}})^{-1}$
$\mathbf{m}_{t_i|t_i} = \mathbf{m}_{t_i|t_{i-1}} + \tau^{-2}\mathbf{R}_{t_i|t_i}\{\tilde{\mathbf{w}}(t_i) - \mathbf{m}_{t_i|t_{i-1}}\}$
end for

that the initial distribution equals the innovation distribution. The spectral Kalman filter has as input the Fourier transform of $\tilde{\mathbf{w}} = (\tilde{\mathbf{w}}(t_1)^{\mathrm{T}}, \ldots, \tilde{\mathbf{w}}(t_T)^{\mathrm{T}})^{\mathrm{T}}$ of $\mathbf{w}$, the diagonal matrix $\mathbf{F}$ given by

$$\begin{aligned}(\mathbf{F})_{1:4,1:4} &= \mathrm{diag}[\exp\{-2\Delta(\mathbf{k}_j^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{k}_j + \zeta)\}],\\ (\mathbf{F})_{5:N,5:N} &= \mathrm{diag}[\exp\{-2\Delta(\mathbf{k}_j^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{k}_j + \zeta)\}\mathbf{1}_2],\end{aligned} \tag{28}$$

and other parameters that characterize the SPDE model. It returns forecast and filter means $\mathbf{m}_{t_i|t_{i-1}}$ and $\mathbf{m}_{t_i|t_i}$ and covariance matrices $\mathbf{R}_{t_i|t_i}$ and $\mathbf{R}_{t_i|t_{i-1}}$, $i = 1, \ldots, T$, respectively, i.e. $\mathbf{m}_{t_i|t_i}$ and $\mathbf{R}_{t_i|t_i}$ are the mean and the covariance matrix of $\boldsymbol{\alpha}(t_i)$ given data up to time t_i $\{\mathbf{w}(t_j)|\, j = 1, \ldots, i\}$. Analogously, $\mathbf{m}_{t_i|t_{i-1}}$ and $\mathbf{R}_{t_i|t_{i-1}}$ 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 $\tilde{\mathbf{Q}}$ and $\mathbf{F}$ are diagonal, the covariance matrices $\mathbf{R}_{t_i|t_i}$ and $\mathbf{R}_{t_i|t_{i-1}}$ are also diagonal. Note that the matrix notation in algorithm 1 is used solely for illustration. In practice, matrix vector products ($\mathbf{G}\mathbf{m}_{t_{i-1}|t_{i-1}}$), matrix multiplications ($\mathbf{R}_{t_{i-1}|t_{i-1}}\mathbf{F}$) and matrix inversions $(\tau^{-2} + \mathbf{R}_{t_i|t_{i-1}})^{-1}$ are not calculated with general purpose algorithms but elementwise since all matrices are diagonal or 2×2 block diagonal. It follows that the computational cost for this algorithm is $O(TN)$.

The derivation of algorithm 1 follows from the classical Kalman filter (see, for example, Künsch (2001)) using $\boldsymbol{\Phi}'\boldsymbol{\Phi} = \mathbf{1}_N$, $\mathbf{G}\mathbf{R}_{t_{i-1}|t_{i-1}}\mathbf{G}^{\mathrm{T}} = \mathbf{R}_{t_{i-1}|t_{i-1}}\mathbf{G}\mathbf{G}^{\mathrm{T}}$, and the fact that $\mathbf{G}\mathbf{G}^{\mathrm{T}} = \mathbf{F}$. The first equation holds true because of the orthonormality of the discrete Fourier transform. The second equation follows from the fact that $\mathbf{G}$ is 2×2 block diagonal and that $\mathbf{R}_{t_{i-1}|t_{i-1}}$ is diagonal with the diagonal entries being equal for each cosine–sine pair. The last equation holds true as shown in what follows. Being obvious for the first four frequencies, we consider the 2×2 diagonal blocks of cosine–sine pairs:

$$\begin{aligned}&(\mathbf{G})_{(2l-5):(2l-4),(2l-5):(2l-4)}(\mathbf{G})^{\mathrm{T}}_{(2l-5):(2l-4),(2l-5):(2l-4)}\\ &= \exp\{-2\Delta(\mathbf{k}_j^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{k}_j + \zeta)\}(\cos(\Delta\boldsymbol{\mu}^{\mathrm{T}}\mathbf{k}_j)\mathbf{1}_2 - \sin(\Delta\boldsymbol{\mu}^{\mathrm{T}}\mathbf{k}_j)\mathbf{J}_2)(\cos(\Delta\boldsymbol{\mu}^{\mathrm{T}}\mathbf{k}_j)\mathbf{1}_2 - \sin(\Delta\boldsymbol{\mu}^{\mathrm{T}}\mathbf{k}_j)\mathbf{J}_2)^{\mathrm{T}}\\ &= \exp\{-2\Delta(\mathbf{k}_j^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{k}_j + \zeta)\}(\cos(\Delta\boldsymbol{\mu}^{\mathrm{T}}\mathbf{k}_j)^2 + \sin(\Delta\boldsymbol{\mu}^{\mathrm{T}}\mathbf{k}_j)^2)\mathbf{1}_2,\end{aligned}$$

$l = 5, \ldots, N/2 + 2$, which equals equation (28). In the last equation we have used $\mathbf{J}_2^{\mathrm{T}} = -\mathbf{J}_2$ and $\mathbf{J}_2^2 = -\mathbf{1}_2$.

On the basis of the Kalman filter, the log-likelihood is calculated as (see, for example, Shumway and Stoffer (2000))

$$l = \sum_{i=1}^{T} \log|\mathbf{R}_{t_i|t_{i-1}} + \tau^2\mathbf{1}_N| + (\tilde{\mathbf{w}}(t_i) - \mathbf{m}_{t_i|t_{i-1}})^{\mathrm{T}}(\mathbf{R}_{t_i|t_{i-1}} + \tau^2\mathbf{1}_N)^{-1}(\tilde{\mathbf{w}}(t_i) - \mathbf{m}_{t_i|t_{i-1}}) + \frac{TN}{2}\log(2\pi). \tag{29}$$

Table 2. Algorithm 2: spectral backward sampling

Input: T, $\mathbf{G}$, $\tilde{\mathbf{Q}}$, $\mathbf{F}$, $\mathbf{m}_{t_i|t_{i-1}}$, $\mathbf{m}_{t_i|t_i}$, $\mathbf{R}_{t_i|t_i}$ and $\mathbf{R}_{t_i|t_{i-1}}$, $i = 1, \ldots, T$
Output: a sample $\boldsymbol{\alpha}^*(t_1), \ldots, \boldsymbol{\alpha}^*(t_T)$ from $[\boldsymbol{\alpha}|\cdot]$

$\boldsymbol{\alpha}^*(t_T) = \mathbf{m}_{t_T|t_T} + \mathbf{R}_{t_T|t_T}^{1/2}\mathbf{n}_T, \mathbf{n}_T \sim N(\mathbf{0}, \mathbf{1}_N)$
for $i = T-1, \ldots, 1$ do
  $\bar{\mathbf{m}}_{t_i} = \mathbf{m}_{t_i|t_i} + \mathbf{R}_{t_i|t_i}\mathbf{R}_{t_i|t_{i-1}}^{-1}\mathbf{G}^{\mathrm{T}}\{\boldsymbol{\alpha}^*(t_{i+1}) - \mathbf{m}_{t_i|t_{i-1}}\}$
  $\bar{\mathbf{R}}_{t_i} = (\tilde{\mathbf{Q}}\mathbf{F} + \mathbf{R}_{t_{i-1}|t_{i-1}}^{-1})^{-1}$
  $\boldsymbol{\alpha}^*(t_i) = \bar{\mathbf{m}}_{t_i} + \bar{\mathbf{R}}_{t_i}^{1/2}\mathbf{n}_i, \mathbf{n}_i \sim N(\mathbf{0}, \mathbf{1}_N)$
end for

Since the forecast covariance matrices $\mathbf{R}_{t_i|t_{i-1}}$ are diagonal, calculation of their determinants and their inverses is trivial, and the computational cost is again $O(TN)$.

In a Bayesian context, the main difficulty consists in simulating from the full conditional of the latent coefficients $[\boldsymbol{\alpha}|\cdot]$. After running the Kalman filter, this can be done with a backward sampling step. Together, these two algorithms are known as FFBS (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994). Again, backward sampling is computationally very efficient in the spectral space with cost being $O(TN)$. Algorithm 2 (Table 2) shows the backward sampling algorithm in the spectral space. The matrices $\bar{\mathbf{R}}_{t_i}$ are diagonal, which makes their Cholesky decomposition trivial.

4.2. Dimension reduction and missing or non-gridded data

If desired, the total computational cost can be additionally alleviated by using a reduced dimensional Fourier basis with $K \ll N$, N being the number of grid points. This means that we include only certain frequencies, typically low frequencies. 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 0. Performing the FFT still requires $O\{TN\log(N)\}$ operations, though.

When the observed data do not lie on a grid or have 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 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. An FFT can then be applied to the augmented data, and the algorithms presented above can be used. Alternatively, as in our application, one can include an incidence matrix $\mathbf{H}$ that relates the process on the grid to the observation locations. Instead of expression (27), the model is then

$$\mathbf{w}(t_{i+1}) = \mathbf{H}\boldsymbol{\Phi}\,\boldsymbol{\alpha}(t_{i+1}) + \boldsymbol{\nu}(t_{i+1}), \qquad \boldsymbol{\nu}(t_{i+1}) \sim N(0, \tau^2\mathbf{1}_N). \tag{30}$$

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

4.3. A Markov chain Monte Carlo algorithm for Bayesian inference

On the basis of the algorithms presented above, there are several possible ways for doing statistical inference. For instance, if one adopts a frequentist paradigm, one can numerically maximize the log-likelihood (29). In what follows, we briefly present how Bayesian inference can be done

by using an MCMC algorithm (see Gilks *et al.* (1996), Robert and Casella (2004) and Brooks *et al.* (2011)). This algorithm is implemented in the R package `spate` (Sigrist *et al.*, 2012) and used in the application in Section 5.

To complete the specification of a Bayesian model, prior distributions for the parameters $\boldsymbol{\theta} = (\rho_0, \sigma^2, \zeta, \rho_1, \gamma, \alpha, \mu_x, \mu_y, \tau^2)^{\mathrm{T}}$ must be chosen. In general, this choice can depend on the specific application. We present choices for priors that are weakly uninformative. On the basis of Gelman (2006), we suggest the use of improper priors for the σ^2 (marginal variance of the innovation) and τ^2 (nugget effect variance) that are uniform on the standard deviation scale σ and τ respectively. Further, the drift parameters μ_x and μ_y have uniform priors on $[-0.5, 0.5]$, ψ (the direction of anisotropy) has a uniform prior on $[0, \pi/2]$ and γ (the degree of anisotropy) has a uniform prior on the log-scale of the interval $[0.1, 10]$. γ is restricted to $[0.1, 10]$ since stronger anisotropy does not seem reasonable. The range parameters of the innovations and the diffusion matrix ρ_0 and ρ_1 respectively as well as the damping parameter ζ are assigned improper, locally uniform priors on $\mathbb{R}_+$.

Our goal is then to simulate from the joint posterior of the unobservables $[\boldsymbol{\theta}, \boldsymbol{\alpha}|\mathbf{w}]$, where $\mathbf{w}$ denotes the set of all observations. Missing data can be accommodated by using a data augmentation approach which results in an additional Gibbs step; see Section 5.3. Since the latent process $\boldsymbol{\xi}$ is the Fourier transform of the coefficients $\boldsymbol{\alpha}$, $\boldsymbol{\xi}(t_i) = \boldsymbol{\Phi}\,\boldsymbol{\alpha}(t_i)$, sampling from the posterior of $\boldsymbol{\alpha}$ is, from a methodological point of view, equivalent to sampling from the posterior of $\boldsymbol{\xi}$. In what follows, 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 $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$. One could also further divide the latent process $\boldsymbol{\alpha}$ in blocks by iteratively sampling $\boldsymbol{\alpha}(t_i)$ at each time point. However $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ can be strongly dependent, which results in slow mixing. This problem is similar to that observed when doing inference for diffusion models; see, for example, Roberts and Stramer (2001) and Golightly and Wilkinson (2008). It is therefore recommendable to sample jointly from $[\boldsymbol{\theta}, \boldsymbol{\alpha}|\mathbf{w}]$ in a Metropolis–Hastings step.

Joint sampling from $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ is done as follows. First, a proposal $(\boldsymbol{\theta}^*, \boldsymbol{\alpha}^*)$ is obtained by sampling $\boldsymbol{\theta}^*$ from a Gaussian distribution with the mean equalling the last value and an adaptively estimated proposal covariance matrix. To be more specific, ρ_0, σ^2, ζ, ρ_1, γ and τ^2 are sampled on a log-scale to ensure that they remain positive. Then, a sample $\boldsymbol{\alpha}^*$ from $[\boldsymbol{\alpha}|\boldsymbol{\theta}^*, \mathbf{w}]$ is obtained by using the FFBS algorithm (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994). It can be shown that the acceptance ratio for the joint proposal is

$$\min\left\{1, \frac{P(\boldsymbol{\theta}^*|\mathbf{w})\,P(\boldsymbol{\theta}^*)\rho_0^*\sigma^{2*}\zeta^*\rho_1^*\gamma^*\tau^{2*}}{P(\boldsymbol{\theta}^{(i)}|\mathbf{w})\,P(\boldsymbol{\theta}^{(i)})\rho_0^{(i)}\sigma^{2(i)}\zeta^{(i)}\rho_1^{(i)}\gamma^{(i)}\tau^{2(i)}}\right\}, \tag{31}$$

where $P(\boldsymbol{\theta}|\mathbf{w})$ denotes the likelihood of $\boldsymbol{\theta}$ given $\mathbf{w}$ and $P(\boldsymbol{\theta})$ the prior, and where $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^{(i)}$ denote the proposal and the last values respectively. The factor $\rho_0\sigma^2\zeta\rho_1\gamma\tau^2$ is included since these parameters are sampled on a log-scale. We see that this acceptance ratio does not depend on the latent process $\boldsymbol{\xi} = \boldsymbol{\Phi}\boldsymbol{\alpha}$. Thus, the parameters $\boldsymbol{\theta}$ are allowed to move faster in their parameter space. The value of the likelihood $P(\boldsymbol{\theta}|\mathbf{w})$ is obtained as a side product of the Kalman filter in the FFBS.

For this random-walk Metropolis step, we suggest the use of an adaptive algorithm (Roberts and Rosenthal, 2009), meaning that the proposal covariance matrices for $\boldsymbol{\theta}$ 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 5), the fixed effects can also be strongly dependent with the random effects $\boldsymbol{\xi}$. This means that it is advisable that the coefficients $\mathbf{b} \in \mathbb{R}^p$ of the potential covariates $\mathbf{x}(t, \mathbf{s}) \in \mathbb{R}^p$ are also sampled together with $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$. This can be done by slightly modifying the above algorithm. First, the regression coefficients $\mathbf{b}^*$ are proposed jointly with $\boldsymbol{\theta}^*$ in a random-walk Metropolis step. Then $\boldsymbol{\alpha}^*$ is sampled from $[\boldsymbol{\alpha}|\boldsymbol{\theta}^*, \mathbf{b}^*, \mathbf{w}]$ analogously by using the FFBS. Finally, in the acceptance ratio (31), $P(\boldsymbol{\theta}|\mathbf{w})$ now must just be replaced by $P(\boldsymbol{\theta}, \mathbf{b}|\mathbf{w})$, which is also a side product of the Kalman filter.

5. Post-processing precipitation forecasts

NWP models are capable of producing predictive fields at spatially and temporally high frequencies. Statistical post-processing, which is the main objective of this application, serves two purposes. First, probabilistic predictions are obtained in cases where only deterministic predictions are available. Further, even if 'probabilistic' forecasts in the 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 post-processing is then to obtain calibrated and sharp predictive distributions (see Gneiting *et al.* (2007a) for a definition of calibration and sharpness). For precipitation, the need for post-processing 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 post-processing 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 a prior climatic distribution (Krzysztofowicz and Maranzano, 2006), neural networks (Ramrez *et al.*, 2005) and binning techniques (Yussouf and Stensrud, 2006). Sloughter *et al.* (2007) proposed a two-stage model to post-process precipitation forecasts. Berrocal *et al.* (2008) extended the model of Sloughter *et al.* (2007) by accounting for spatial correlation. Kleiber *et al.* (2011) presented a similar model that includes ensemble predictions and accounts for spatial correlation.

Except for the last two references, spatial correlation is typically not modelled in post-processing precipitation forecasts, and none of the aforementioned models explicitly accounts for spatiotemporal dependences. However, for temporally and spatially highly resolved data, it is necessary to account for correlation in space and time. First, spatiotemporal 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, including interactions between space and time. The SPDE approach allows for such interactions, as do other approaches which use scientifically based physical models (Wikle and Hooten, 2010).

5.1. Data

The goal is to post-process 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 Consortium for Small-scale Modelling (see, for example, Steppeler *et al.* (2003)). The NWP model produces deterministic forecasts once a day starting at 0:00 Universal Time Co-ordinated (UTC). Predictions are made for eight consecutive time periods corresponding to 24 h ahead. In what follows, let $y_F(t, \mathbf{s})$ denote the forecast of the rainfall sum from time $t-1$ to t at site s made at 0:00 UTC of the same day. We consider a rectangular region in northern Switzerland shown in

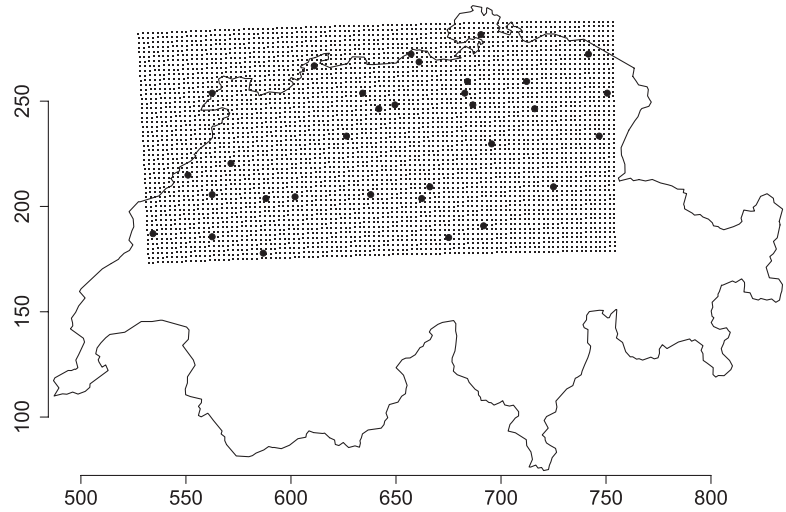

Fig. 5. Locations of grid points at which predictions are obtained (·, 50 × 100 grid) and observations (•): both axes are in kilometres using the Swiss co-ordinate system (CH1903)

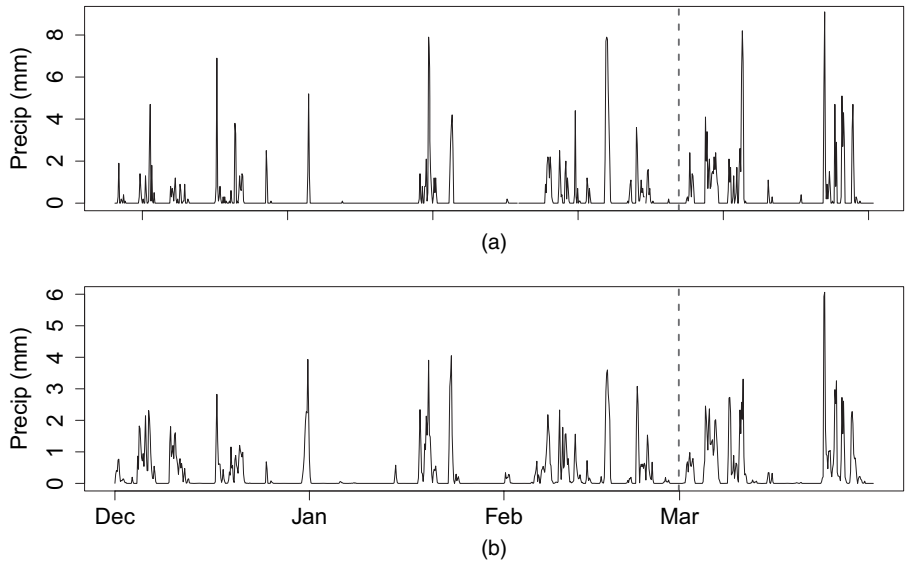

Fig. 6. Precipitation *versus* time, (a) for one station (Waedenswil) and (b) averaged over all stations

Fig. 5. The grid at which predictions are made is of size 50×100. Precipitation is observed at 32 stations over northern Switzerland. Fig. 5 also shows the locations of the observation stations. In the post-processing model, the NWP forecasts are used as covariates in a regression term; see expression (33). We use data for 3-hourly rainfall amounts from the beginning of December 2008 till the end of March 2009. To illustrate the observed data, in Fig. 6, observed precipitation at one station and the equally weighted areal average precipitation are plotted against time. We shall use the first three months containing 720 time points for fitting, and the last month is left aside for evaluation.

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.

5.2. Precipitation model for post-processing

The model that is presented in what follows is a Bayesian hierarchical model. It uses the SPDE-based spatiotemporal Gaussian process $\xi(t,\mathbf{s})$ that was presented in Section 3 at the process level. At the data stage, a mixture model adapted to the nature of precipitation is used. A characteristic feature of precipitation is that its distribution consists of a discrete component, indicating the occurrence of precipitation, and a continuous component, determining the amount (see Fig. 6). As a consequence, there are two basic statistical modelling 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, for example, Sigrist *et al.* (2012) for a more extensive overview of precipitation models and for further details on the data model that is used below. Originally, the approach that is presented in what follows goes back to Tobin (1958) who analysed household expenditure on durable goods. For modelling 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. Sansó and Guenni (1999) developed Bayesian methods for the spatiotemporal analysis of rainfall by using this skewed tobit model, but in contrast with our application they did not explicitly account for temporal correlation and they used a much smaller spatial grid.

We denote the cumulative rainfall from time $t-1$ to t at site $\mathbf{s}\in\mathbb{R}^2$ by $y(t,\mathbf{s})$ and assume that it depends on a latent Gaussian variable $w(t,\mathbf{s})$ through

$$y(t,\mathbf{s}) = \begin{cases} 0, & \text{if } w(t,\mathbf{s}) \leqslant 0, \\ w(t,\mathbf{s})^{\lambda}, & \text{if } w(t,\mathbf{s}) > 0, \end{cases} \tag{32}$$

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 $w(t,\mathbf{s})$ is interpreted as a precipitation potential.

The mean of the Gaussian process $w(t,\mathbf{s})$ is assumed to depend linearly on spatiotemporal covariates $\mathbf{x}(t,\mathbf{s})\in\mathbb{R}^k$. As shown below, this mean term basically consists of the NWP forecasts. Variation that is not explained by the linear term is modelled by using the Gaussian process $\xi(t,\mathbf{s})$ and the unstructured term $\nu(t,\mathbf{s})$ for microscale variability and measurement errors. The spatiotemporal process $\xi(t,\mathbf{s})$ 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 the post-processed forecast is probabilistic and its distribution sharp and calibrated.

To be more specific concerning the covariates, similarly to what appears in Berrocal *et al.* (2008), we include a transformed variable $y_{\mathrm{F}}(t,\mathbf{s})^{1/\tilde{\lambda}}$ and an indicator variable $\mathbb{1}_{\{y_{\mathrm{F}}(t,\mathbf{s})=0\}}$ which equals 1 if $y_{\mathrm{F}}(t,\mathbf{s})=0$ and 0 otherwise. $\tilde{\lambda}$ is determined by fitting the transformed tobit model as in expression (32) to the marginal distribution of the rain data ignoring any spatiotemporal correlation. In doing so, we obtain $\tilde{\lambda}\approx 1.4$. $y_{\mathrm{F}}(t,\mathbf{s})^{1/\lambda}$ is centred near zero by subtracting its overall mean $\bar{y}_{\mathrm{F}}^{1/\lambda}$ to reduce posterior correlations. Thus,

$$w(t,\mathbf{s}) = b_1\{y_{\mathrm{F}}(t,\mathbf{s})^{1/\tilde{\lambda}} - \bar{y}_{\mathrm{F}}^{1/\tilde{\lambda}}\} + b_2\mathbb{1}_{\{y_{\mathrm{F}}(t,\mathbf{s})=0\}} + \xi(t,\mathbf{s}) + \nu(t,\mathbf{s}). \tag{33}$$

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

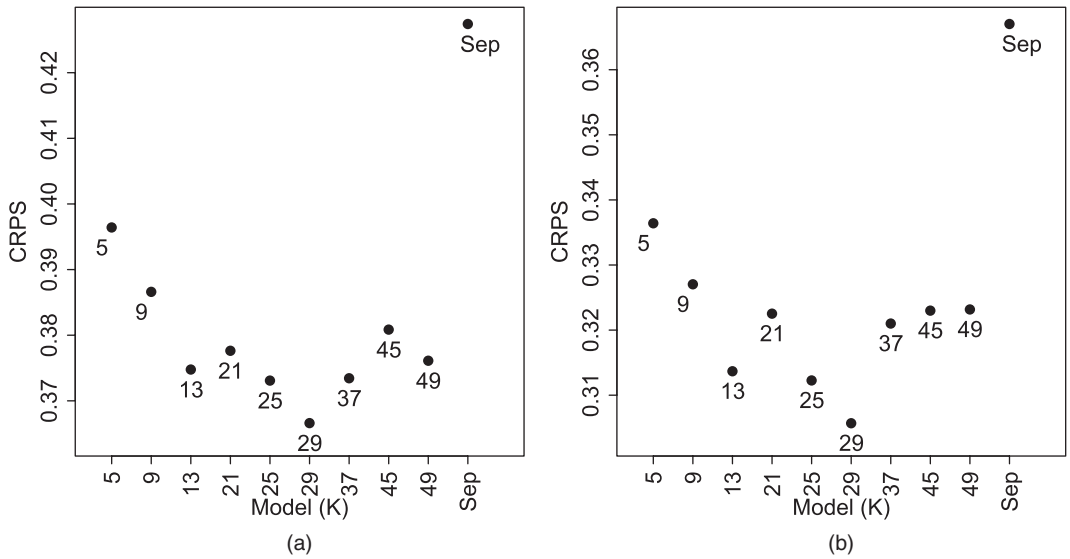

Fig. 7. Comparison of various statistical models by using the continuous ranked probability score (in millimetres (K denotes the number of basis functions used in the model; 'Sep' denotes the separable model with $K = 29$ Fourier terms): (a) continuous ranked probability scores of station-specific forecasts; (b) continuous ranked probability scores of areal forecasts

algorithm that is 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 $\boldsymbol{\alpha}$. Further, unidentifiability is unlikely to be a problem in these cases.

Concerning the spatiotemporal process $\xi(t, \mathbf{s})$, 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 with the spatial domain, and padding is therefore used 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 $w(t, \mathbf{s})$ can only be modelled on the 50×100 grid where the covariates are available. To cope with this we use an incidence matrix $\mathbf{H}$ as in expression (30) to relate the process at the 200×200 grid to the observation stations. As argued in Section 4.2, this then requires that we use a reduced dimensional Fourier expansion, i.e., instead of using $N = 200^2$ basis functions, we use only $K \ll N$ 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 is confirmed by our analysis; Fig. 7.

Concerning prior distributions, for $\boldsymbol{\theta} = (\rho_0, \sigma^2, \zeta, \rho_1, \gamma, \psi, \mu_x, \mu_y, \tau^2)^{\mathrm{T}}$, we use the priors that were presented in Section 4.3. The parameters $\mathbf{b}$ and λ, which are not included in $\boldsymbol{\theta}$, have improper, locally uniform priors on $\mathbb{R}$ and $\mathbb{R}_+$ respectively. In summary,

$$P(\mathbf{b}, \lambda, \boldsymbol{\theta}) \propto \frac{1}{\gamma \sqrt{\sigma^2} \sqrt{\tau^2}} \mathbb{1}_{\{-0.5 \leqslant \mu_x, \mu_y \leqslant 0.5\}} \mathbb{1}_{\{0 \leqslant \psi \leqslant \pi/2\}} \mathbb{1}_{\{\lambda, \rho_0, \rho_1, \zeta, \sigma^2, \tau^2 \geqslant 0\}} \mathbb{1}_{\{0.1 \leqslant \gamma \leqslant 10\}}.$$

In addition, concerning $\boldsymbol{\alpha}(0)$, we choose to use the innovation distribution that is specified in expression (6) as the initial distribution.

5.3. Fitting

MCMC sampling is used to sample from the posterior distribution $[\mathbf{b}, \lambda, \boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{w} | \mathbf{y}]$, where $\mathbf{y}$

denotes the set of all observations. We use what Neal and Roberts (2006) called a Metropolis-within-Gibbs algorithm which alternates between blocked Gibbs (Gelfand and Smith, 1990) and Metropolis (Metropolis *et al.*, 1953; Hastings, 1970) sampling steps.

We use the Metropolis–Hastings algorithm that was presented in Section 4.3 with the coefficients $\mathbf{b}$ being sampled jointly with $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$. Owing to the non-Gaussian data model, additional Metropolis and Gibbs steps are required for λ and for those points of $\mathbf{w}$ where the observed rainfall amount is 0 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 $\mathbf{w}^{[0]}$ the values of $\mathbf{w}$ at those points where the observed rainfall is 0: $y(t, \mathbf{s}) = 0$. Analogously, we define $\mathbf{w}^{[m]}$ and $\mathbf{w}^{[+]}$ for the missing values and the values where a positive rainfall amount is observed, $y(t, \mathbf{s}) > 0$, respectively. The full conditionals of the censored $\mathbf{w}^{[0]}$ and missing points $\mathbf{w}^{[m]}$ are truncated and regular one-dimensional Gaussian distributions respectively. Sampling from them is done in Gibbs steps. The transformation parameter λ is sampled by using a random-walk Metropolis step. If a new value is accepted, $\mathbf{w}^{[+]}$ needs to be updated by using the deterministic relationship $w(t, \mathbf{s}) = y(t, \mathbf{s})^{1/\lambda}$ due to expression (32). From these Gibbs and Metropolis steps, we obtain $\mathbf{w}$ consisting of simulated and transformed observed data. In the second part of the algorithm, we sample $\mathbf{b}$, $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ jointly from $[\mathbf{b}, \boldsymbol{\theta}, \boldsymbol{\alpha} | \mathbf{w}]$ by using the algorithm that was presented in Section 4.3, where $\mathbf{w}$ acts as if it were the observed data. After a burn-in of 5000 iterations, we use 100000 samples from the Markov chain to characterize the posterior distribution. Convergence is monitored by inspecting trace plots.

5.4. Model selection and results

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

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

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} \{F(x) - \mathbb{1}_{\{y \leqslant x\}}\}^2 \, \mathrm{d}x, \tag{34}$$

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

$$\frac{1}{m} \sum_{i=1}^{m} |y^{(i)} - y| - \frac{1}{2m^2} \sum_{i,j=1}^{m} |y^{(i)} - y^{(j)}|. \tag{35}$$

Ideally, we would run the full MCMC algorithm at each time point $t \geqslant 720$, 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 $\boldsymbol{\theta}$, $\mathbf{b}$ and λ given $\mathbf{y}_{1:t} = \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$ is the same for all $t \geqslant 720$, i.e. we neglect the additional information that the observations in March provide 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. The 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 are used to obtain the posterior distribution. But, since all our data lie 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 by using a temporal evolution.

For each time point $t \geqslant 720$, we make up to eight-steps-ahead forecasts corresponding to 24 h, i.e. we sample from the predictive distribution of $\mathbf{y}^*_{t+k}$, $k = 1, \ldots, 8$, given $\mathbf{y}_{1:t} = \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$.

In Fig. 7, the average CRPS of the pointwise predictions and the areal predictions are shown for the various statistical models. In Fig. 7(a) 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 in lower predictive performance. We interpret this result 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 are accounted for by the forecast y_F. For comparison, we also fit a separable model which is obtained by setting $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma}^{-1} = \mathbf{0}_{2,2}$. 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. On the basis of these findings, we decided to use the model with 29 cosine and sine functions.

Table 3 shows posterior medians as well as 95% credible intervals for the various parameters. Note that the range parameters ρ_0 and ρ_1 as well as the drift parameters μ_x and μ_y have been transformed back from the unit [0, 1] scale to the original kilometres scale. The posterior median of the variance σ^2 of the innovations of the spatiotemporal process is around 0.8. Compared with this, the nugget variance, about 0.3, is smaller. For the innovation range parameter ρ_0, we obtain a value of about 25 km. And the range parameter ρ_1 that controls the amount of diffusion or, in other words, the amount of spatiotemporal interaction, is approximately 49 km. With γ

Table 3. Posterior medians and 95% credible intervals for the SPDE-based spatiotemporal model presented in Section 3 with $K = 29$ Fourier terms

Parameter	*Median*	*2.5% value*	*97.5% value*
ρ_0	25.4	18.8	32.4
σ^2	0.838	0.727	0.994
ζ	0.00655	0.000395	0.0156
ρ_1	48.8	42.1	57.1
γ	4.33	3.34	6.01
ψ	0.557	0.49	0.617
μ_x	6.73	0.688	12.9
μ_y	−4.19	−8.55	−0.435
τ^2	0.307	0.288	0.327
b_1	0.448	0.414	0.481
b_2	−0.422	−0.5	−0.344
λ	1.67	1.64	1.7

Table 4. Comparison of the NWP model and statistically post-processed forecasts by using the mean absolute error†

	Mean absolute errors (mm) for the following forecasts:		
	Post processd	*NWP*	*Static*
Stationwise	0.359	0.485	0.594
Areal	0.303	0.387	0.489

†'Static' denotes the constant forecast obtained by using the most recently observed data.

and ψ being around 4 and 0.6 respectively, we observe anisotropy in the south-west–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 ζ has a posterior median of about 0.01.

Next, we compare the performance of the post-processed forecasts with those 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 stations (the results are not reported). Since the NWP produces eight-step-ahead predictions once a day, we consider only statistical forecasts starting at 0:00 UTC. This is in contrast with the above comparison of the different statistical models for which eight-step-ahead predictions were made at all time points and not just once for each day. We use the mean absolute error for evaluating the NWP forecasts. To be consistent, we also generate point forecasts from the statistical predictive distributions by using medians and then calculate the mean absolute error for these point forecasts. In Table 4, 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 post-processed forecasts clearly perform better than the raw NWP forecasts. In addition, the post-processed forecasts have the advantage that they provide probabilistic forecasts quantifying prediction uncertainty.

The statistical model produces a joint spatiotemporal predictive distribution that is spatially highly resolved. To illustrate the use of the model, we show several quantities in Fig. 8. We consider the time point $t = 760$ and calculate predictive distributions over the next 24 h. Predicted fields for the period $t = 761, \ldots, 768$ from the NWP are shown in Fig. 8(a). In Fig. 8(b) are pointwise medians obtained from the statistical forecasts. This is a period during which the NWP predicts too much rainfall compared with the observed data (the results are not shown). Fig. 8 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 (which are not shown here), that can be easily obtained, include probabilities of precipitation occurrence or various quantiles of the distribution.

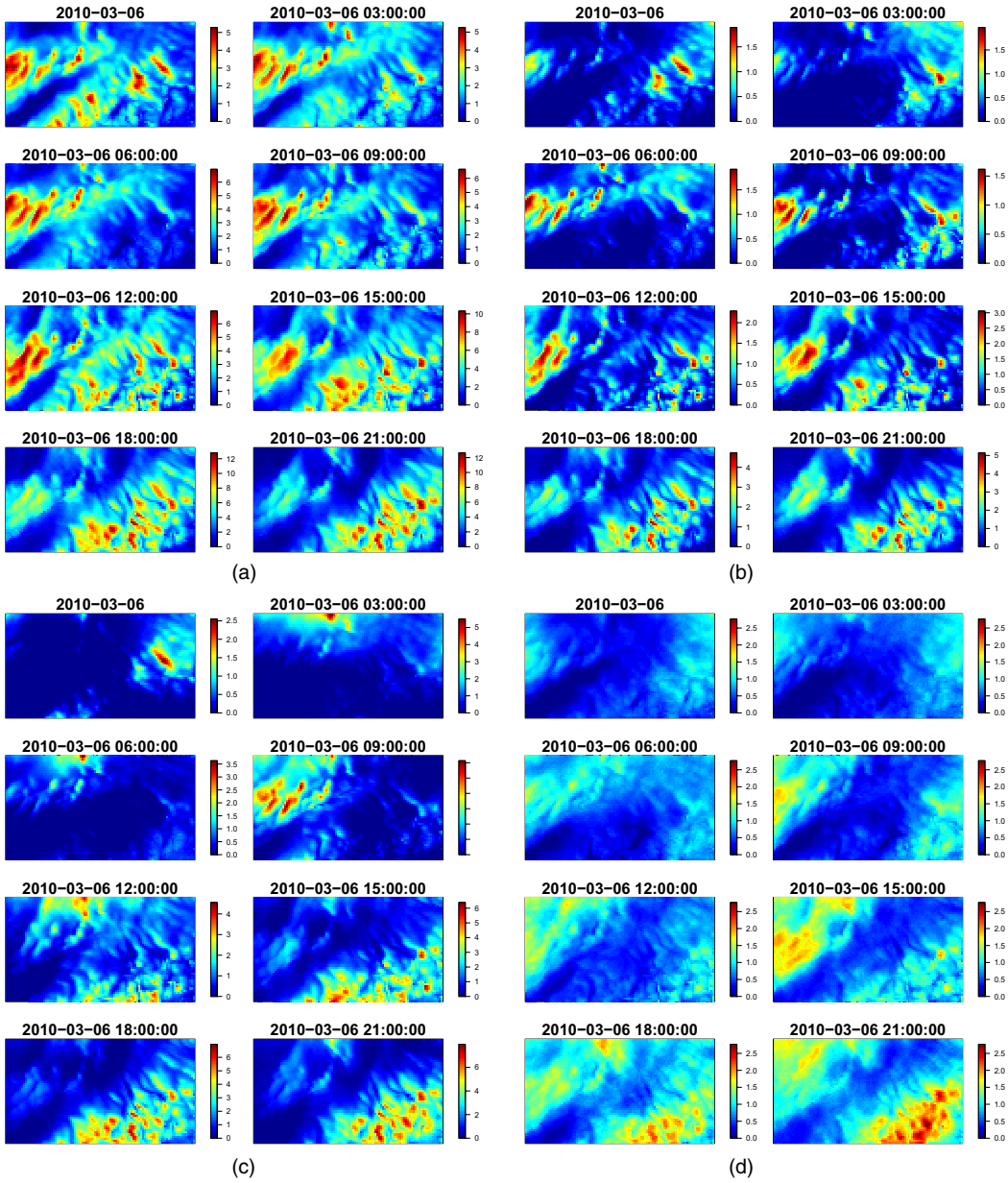

Fig. 8. Illustration of post-processed spatiotemporal precipitation fields for the period $t = 761,\ldots,768$ (all quantities are in millimetres; note that the scales are different in different figures): (a) NWP forecasts; (b) pointwise medians of the predictive distribution; (c) one sample from the predictive distribution; (d) differences between the third quartile and the median of the predictive distribution

6. Conclusion

We present a spatiotemporal 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 by using Fourier functions,

and we have given a bound on the precision of the approximate solution. In the spectral space, we can use computationally efficient statistical algorithms whose computational costs grow linearly with the dimension, the total computational costs being dominated by the FFT. The space–time Gaussian process that is defined through the advection–diffusion SPDE has a non-separable covariance structure and can be physically motivated. The model is applied to post-processing of precipitation forecasts for northern Switzerland. The post-processed forecasts clearly outperform the raw NWP predictions. In addition, they have the advantage that they quantify prediction uncertainty.

In our analysis, we considered cumulative rainfall over 3 h, both in the NWP forecasts and in the station data. It would be interesting to formulate a model which can describe different accumulation periods in a coherent way and is still computationally feasible. Another 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. Since the operators of the SPDE are local, we can define the SPDE on general manifolds and, in particular, on the sphere (see, for example, Lindgren *et al.* (2011)). Future research will show to what extent spectral methods can still be used in practice.

Acknowledgements

We are grateful to Vanessa Stauch from MeteoSchweiz for providing the data and for inspiring discussions. In addition, we thank Peter Guttorp for interesting comments and discussions and two referees for helpful comments and suggestions.

References

Abramowitz, M. and Stegun, I. A. (1964) *Handbook of Mathematical Functions*. New York: Dover Publications.

Anderson, B. D. O. and Moore, J. B. (1979) *Optimal Filtering* (eds B. D. O. Anderson and J. B. Moore). Englewood Cliffs: Prentice Hall.

Antolik, M. S. (2000) An overview of the national weather service's centralized statistical quantitative precipitation forecasts. *J. Hydrol.*, **239**, 306–337.

Applequist, S., Gahrs, G. E., Pfeffer, R. L. and Niu, X.-F. (2002) Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Weath. Forecast.*, **17**, 783–799.

Aune, E. and Simpson, D. (2012) The use of systems of stochastic PDEs as priors for multivariate models with discrete structures. *Preprint arXiv:1208.1717*.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman and Hall–CRC.

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008) Gaussian predictive process models for large spatial data sets. *J. R. Statist. Soc.* B, **70**, 825–848.

Bardossy, A. and Plate, E. (1992) Space-time model for daily rainfall using atmospheric circulation patterns. *Wat. Resour. Res.*, **28**, 1247–1259.

Bell, T. (1987) A space-time stochastic model of rainfall for satellite remote-sensing studies. *J. Geophys. Res.*, **92**, 9631–9643.

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

Borgman, L., Taheri, M. and Hagan, R. (1984) Three-dimensional frequency-domain simulations of geological variables. In *Geostatistics for Natural Resources Characterization* (ed. G. Verly), pp. 517–541. Boston: Reidel.

Bremnes, B. J. (2004) Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mnthly Weath. Rev.*, **132**, 338–347.

Bronson, R. and Costa, G. B. (2007) *Linear Algebra: an Introduction*. New York: Academic Press.

Brooks, S., Gelman, A., Jones, G. and Meng, X. (2011) *Handbook of Markov Chain Monte Carlo*. Boca Raton: Chapman and Hall–CRC.

Brown, P. E., Kåresen, K. F., Roberts, G. O. and Tonellato, S. (2000) Blur-generated non-separable space–time models. *J. R. Statist. Soc.* B, **62**, 847–860.

Cameletti, M., Lindgren, F., Simpson, D. and Rue, H. (2013) Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *Adv. Statist. Anal.*, **79**, 109–131.
Carter, C. K. and Kohn, R. (1994) On Gibbs sampling for state space models. *Biometrika*, **81**, 541–553.
Coe, R. and Stern, R. (1982) Fitting models to daily rainfall data. *J. Appl. Meteorol.*, **21**, 1024–1031.
Cooley, J. W. and Tukey, J. W. (1965) An algorithm for the machine calculation of complex Fourier series. *Math. Computn*, **19**, 297–301.
Cramér, H. and Leadbetter, M. R. (1967) *Stationary and Related Stochastic Processes: Sample Function Properties and Their Applications*. New York: Wiley.
Cressie, N. and Huang, H.-C. (1999) Classes of nonseparable, spatio-temporal stationary covariance functions. *J. Am. Statist. Ass.*, **94**, 1330–1340.
Cressie, N. and Johannesson, G. (2008) Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc.* B, **70**, 209–226.
Cressie, N. and Wikle, C. K. (2011) *Statistics for Spatio-temporal Data*. Hoboken: Wiley.
Duan, J. A., Gelfand, A. E. and Sirmans, C. (2009) Modeling space-time data using stochastic differential equations. *Baysn Anal.*, **4**, 733–758.
Dudgeon, D. E. and Mersereau, R. M. (1984) *Multidimensional Digital Signal Processing*. Englewood Cliffs: Prentice Hall.
Eidsvik, J., Shaby, B., Reich, B., Wheeler, M. and Niemi, J. (2012) Estimation and prediction in spatial models with block composite likelihoods using parallel computing. *Technical Report*. Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim. (Available from `http://www.math.ntnu.no/~joeid/ESRWN.pdf`.)
Folland, G. B. (1992) *Fourier Analysis and Its Applications*. Pacific Grove: Wadsworth and Brooks–Cole.
Friederichs, P. and Hense, A. (2007) Statistical downscaling of extreme precipitation events using censored quantile regression. *Mnthly Weath. Rev.*, **135**, 2365–2378.
Frühwirth-Schnatter, S. (1994) Data augmentation and dynamic linear models. *J. Time Ser. Anal.*, **15**, 183–202.
Fuentes, M. (2007) Approximate likelihood for large irregularly spaced spatial data. *J. Am. Statist. Ass.*, **102**, 321–331.
Furrer, R., Genton, M. G. and Nychka, D. (2006) Covariance tapering for interpolation of large spatial datasets. *J. Computnl Graph. Statist.*, **15**, 502–523.
Gelfand, A. E., Banerjee, S. and Gamerman, D. (2005) Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics*, **16**, 465–479.
Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.
Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Baysn Anal.*, **1**, 1–19.
Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds) (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.
Gneiting, T. (2002) Nonseparable, stationary covariance functions for space-time data. *J. Am. Statist. Ass.*, **97**, 590–600.
Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007a) Probabilistic forecasts, calibration and sharpness. *J. R. Statist. Soc.* B, **67**, 243–268.
Gneiting, T., Genton, M. G. and Guttorp, P. (2007b) Geostatistical space-time models, stationarity, separability and full symmetry. In *Modelling Longitudinal and Spatially Correlated Data* (eds B. Finkenstädt, L. Held and V. Isham), pp. 151–175. Boca Raton: Chapman and Hall–CRC.
Gneiting, T. and Raftery, A. E. (2005) Weather forecasting with ensemble methods. *Science*, **310**, 248–249.
Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist. Ass.*, **102**, 359–378.
Golightly, A. and Wilkinson, D. (2008) Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computnl Statist. Data Anal.*, **52**, 1674–1693.
Gottlieb, D. and Orszag, S. A. (1977) *Numerical Analysis of Spectral Methods: Theory and Applications*. Philadelphia: Society for Industrial and Applied Mathematics.
Haberman, R. (2004) *Applied Partial Differential Equations: with Fourier Series and Boundary Value Problems*. Englewood Cliffs: Pearson Prentice Hall.
Hamill, T. M. and Colucci, S. J. (1997) Verification of Eta RSM short-range ensemble forecasts. *Mnthly Weath. Rev.*, **125**, 1312–1327.
Hamill, T. M., Whitaker, J. S. and Wei, X. (2004) Ensemble reforecasting: improving medium-range forecast skill using retrospective forecasts. *Mnthly Weath. Rev.*, **132**, 1434–1447.
Handcock, M. S. and Stein, M. L. (1993) A Bayesian analysis of kriging. *Technometrics*, **35**, 403–410.
Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.
Heine, V. (1955) Models for two-dimensional stationary stochastic processes. *Biometrika*, **42**, 170–178.
Hu, X., Simpson, D., Lindgren, F. and Rue, H. (2013) Multivariate gaussian random fields using systems of stochastic partial differential equations. *Preprint arXiv:1307.1379*.

Huang, H.-C. and Hsu, N.-J. (2004) Modeling transport effects on ground-level ozone using a non-stationary space-time model. *Environmetrics*, **15**, 251–268.
Hutchinson, M. (1995) Stochastic space-time weather models from ground-based data. *Agric. Forst Meteorol.*, **73**, 237–264.
Johannesson, G., Cressie, N. and Huang, H.-C. (2007) Dynamic multi-resolution spatial models. *Environ. Ecol. Statist.*, **14**, 5–25.
Jones, R. and Zhang, Y. (1997) Models for continuous stationary space-time processes. In *Modelling Longitudinal and Spatially Correlated Data* (eds T. Gregoire, D. R. Brillinger, P. J. Diggle, E. Russek-Cohen, W. G. Warren and R. Wolfinger), pp. 289–298. New York: Springer.
Kleiber, W., Raftery, A. E. and Gneiting, T. (2011) Geostatistical model averaging for locally calibrated probabilistic quantitative precipitation forecasting. *J. Am. Statist. Ass.*, **106**, 1291–1303.
Krzysztofowicz, R. and Maranzano, C. J. (2006) Bayesian processor of output for probabilistic quantitative precipitation forecasting. *Technical Report*. Department of Systems Engineering and Department of Statistics, University of Virginia, Charlottesville.
Künsch, H. R. (2001) State space and hidden Markov models. In *Complex Stochastic Systems*, pp. 109–173. Boca Raton: Chapman and Hall–CRC.
Lindgren, F., Rue, H. and Lindström, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). *J. R. Statist. Soc.* B, **73**, 423–498.
Ma, C. (2003) Families of spatio-temporal stationary covariance models. *J. Statist. Planng Inf.*, **116**, 489–501.
Malmberg, A., Arellano, A., Edwards, D. P., Flyer, N., Nychka, D. and Wikle, C. (2008) Interpolating fields of carbon monoxide data using a hybrid statistical-physical model. *Ann. Appl. Statist.*, **2**, 1231–1248.
Matheson, J. E. and Winkler, R. L. (1976) Scoring rules for continuous probability distributions. *Mangmnt Sci.*, **22**, 1087–1096.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.
Neal, P. and Roberts, G. (2006) Optimal scaling for partially updating MCMC algorithms. *Ann. Appl. Probab.*, **16**, 475–515.
Nychka, D., Wikle, C. and Royle, J. A. (2002) Multiresolution models for nonstationary spatial covariance functions. *Statist. Modllng*, **2**, 315–331.
Paciorek, C. J. (2007) Bayesian smoothing with Gaussian processes using Fourier basis functions in the spectralGP package. *J. Statist. Softwr.*, **19**, 1–38.
Paciorek, C. J. and Schervish, M. J. (2006) Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, **17**, 483–506.
Palmer, T. N. (2002) The economic value of ensemble forecasts as a tool for risk assessment: from days to decades. *Q. J. R. Meteorol. Soc.*, **128**, 747–774.
Pedlosky, J. (1987) *Geophysical Fluid Dynamics*. New York: Springer.
Ramrez, M. C. V., de Campos Velho, H. F. and Ferreira, N. J. (2005) Artificial neural network technique for rainfall forecasting applied to the Sao Paulo region. *J. Hydrol.*, **301**, 146–162.
Robert, C. P. and Casella, G. (2004) *Monte Carlo Statistical Methods*, 2nd edn. New York: Springer.
Roberts, G. O. and Rosenthal, J. S. (2001) Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.*, **16**, 351–367.
Roberts, G. O. and Rosenthal, J. S. (2009) Examples of adaptive MCMC. *J. Computnl Graph. Statist.*, **18**, 349–367.
Roberts, G. O. and Stramer, O. (2001) On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika*, **88**, 603–621.
Royle, J. A. and Wikle, C. K. (2005) Efficient statistical mapping of avian count data. *Environ. Ecol. Statist.*, **12**, 225–243.
Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields.* Boca Raton: Chapman and Hall–CRC.
Rue, H. and Tjelmeland, H. (2002) Fitting Gaussian Markov random fields to Gaussian fields. *Scand. J. Statist.*, **29**, 31–49.
Sampson, P. D. and Guttorp, P. (1992) Nonparametric estimation of nonstationary spatial covariance structure. *J. Am. Statist. Ass.*, **87**, 108–119.
Sansó, B. and Guenni, L. (1999) Venezuelan rainfall data analysed by using a Bayesian space–time model. *Appl. Statist.*, **48**, 345–362.
Sansó, B. and Guenni, L. (2004) A Bayesian approach to compare observed rainfall data to deterministic simulations. *Environmetrics*, **15**, 597–612.
Shumway, R. H. and Stoffer, D. S. (2000) *Time Series Analysis and Its Applications*. New York: Springer.
Sigrist, F., Künsch, H. R. and Stahel, W. A. (2012) A dynamic nonstationary spatio-temporal model for short term prediction of precipitation. *Ann. Appl. Statist.*, **6**, 1452–1477.
Sigrist, F., Künsch, H. R. and Stahel, W. A. (2012) spate: an R package for statistical modeling with SPDE based spatio-temporal Gaussian processes. *Technical Report.* Seminar für Statistik, Eidgenössiche Technische Hochschule Zürich, Zurich. (Available from `http://stat.ethz.ch/people/sigrist/spate.pdf`.)

Simpson, D., Lindgren, F. and Rue, H. (2012) In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics*, **23**, 65–74.
Sloughter, J. M., Raftery, A. E., Gneiting, T. and Fraley, C. (2007) Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mnthly Weath. Rev.*, **135**, 3209–3220.
Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 3–23.
Solna, K. and Switzer, P. (1996) Time trend estimation for a geographic region. *J. Am. Statist. Ass.*, **91**, 577–589.
Stein, M. L. (1999) *Interpolation of Spatial Data, Some Theory for Kriging*. New York: Springer.
Stein, M. L. (2005) Space-time covariance functions. *J. Am. Statist. Ass.*, **100**, 310–321.
Stein, M. L. (2008) A modeling approach for large spatial datasets. *J. Kor. Statist. Soc.*, **37**, 3–10.
Stein, M. L., Chi, Z. and Welty, L. J. (2004) Approximating likelihoods for large spatial data sets. *J. R. Statist. Soc.* B, **66**, 275–296.
Stensrud, D. J. and Yussouf, N. (2007) Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecasting system. *Weath. Forecast.*, **22**, 2–17.
Steppeler, J., Doms, G., Schättler, U., Bitzer, H. W., Gassmann, A., Damrath, U. and Gregoric, G. (2003) Meso-gamma scale forecasts using the nonhydrostatic model LM. *Meteorol. Atmosph. Phys.*, **82**, 75–96.
Stidd, C. K. (1973) Estimating the precipitation climate. *Wat. Resour. Res.*, **9**, 1235–1241.
Storvik, G., Frigessi, A. and Hirst, D. (2002) Stationary space-time Gaussian fields and their time autoregressive representation. *Statist. Modllng*, **2**, 139–161.
Stroud, J. R., Stein, M. L., Lesht, B. M., Schwab, D. J. and Beletsky, D. (2010) An ensemble Kalman filter and smoother for satellite data assimilation. *J. Am. Statist. Ass.*, **105**, 978–990.
Tobin, J. (1958) Estimation of relationships for limited dependent variables. *Econometrica*, **26**, 24–36.
Vecchia, A. V. (1988) Estimation and model identification for continuous spatial processes. *J. R. Statist. Soc.* B, **50**, 297–312.
Vivar, J. C. and Ferreira, M. A. R. (2009) Spatiotemporal models for gaussian areal data. *J. Computnl Graph. Statist.*, **18**, 658–674.
Whittle, P. (1954) On stationary processes in the plane. *Biometrika*, **41**, 434–449.
Whittle, P. (1962) Topographic correlation, power-law covariance functions, and diffusion. *Biometrika*, **49**, 305–314.
Whittle, P. (1963) Stochastic processes in several dimensions. *Bull. Inst. Int. Statist.*, **40**, 974–994.
Wikle, C. K. (2002) A kernel-based spectral model for non-gaussian spatio-temporal processes. *Statist. Modllng*, **2**, 299–314.
Wikle, C. K. (2003) Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*, **84**, 1382–1394.
Wikle, C. (2010) Low-rank representations for spatial processes. In *Handbook of Spatial Statistics* (eds A. E. Gelfand, P. Diggle, P. Guttorp and M. Fuentes), pp. 107–118. Boca Raton: Chapman and Hall–CRC.
Wikle, C. K., Berliner, L. M. and Cressie, N. (1998) Hierarchical Bayesian space-time models. *Environ. Ecol. Statist.*, **5**, 117–154.
Wikle, C. K. and Cressie, N. (1999) A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, **86**, 815–829.
Wikle, C. and Hooten, M. (2010) A general science-based framework for dynamical spatiotemporal models. *Test*, **19**, 417–451.
Wikle, C. K., Milliff, R. F., Nychka, D. and Berliner, L. M. (2001) Spatiotemporal hierarchical bayesian modeling: tropical ocean surface winds. *J. Am. Statist. Ass.*, **96**, 382–397.
Wilks, D. (1990) Maximum likelihood estimation for the gamma distribution using data containing zeros. *J. Clim.*, **3**, 1495–1501.
Wilks, D. (1999) Multisite downscaling of daily precipitation with a stochastic weather generator. *Clim. Res.*, **11**, 125–136.
Xu, K. and Wikle, C. K. (2007) Estimation of parameterized spatio-temporal dynamic models. *J. Statist. Planng Inf.*, **137**, 567–588.
Xu, K., Wikle, C. K. and Fox, N. I. (2005) A kernel-based spatio-temporal dynamical model for nowcasting weather radar reflectivities. *J. Am. Statist. Ass.*, **100**, 1133–1144.
Yue, Y. R., Simpson, D., Lindgren, F. and Rue, H. (2012) Bayesian adaptive smoothing spline using stochastic differential equations. *Preprint arXiv:1209.2013*.
Yussouf, N. and Stensrud, D. J. (2006) Prediction of near-surface variables at independent locations from a bias-corrected ensemble forecasting system. *Mnthly Weath. Rev.*, **134**, 3415–3424.
Zheng, Y. and Aukema, B. H. (2010) Hierarchical dynamic modeling of outbreaks of mountain pine beetle using partial differential equations. *Environmetrics*, **21**, 801–816.