Infinite-Dimensional Kalman Filtering Approach to Spatio-Temporal Gaussian Process Regression

Simo Särkkä
Department of Biomedical Engineering
and Computational Science
Aalto University
00076 AALTO, Finland
simo.sarkka@aalto.fi

Jouni Hartikainen
Department of Biomedical Engineering
and Computational Science
Aalto University
00076 AALTO, Finland
jouni.hartikainen@aalto.fi

Abstract

We show how spatio-temporal Gaussian process (GP) regression problems (or the equivalent Kriging problems) can be formulated as infinite-dimensional Kalman filtering and Rauch-Tung-Striebel (RTS) smoothing problems, and present a procedure for converting spatio-temporal covariance functions into infinite-dimensional stochastic differential equations (SDEs). The resulting infinite-dimensional SDEs belong to the class of stochastic pseudo-differential equations and can be numerically treated using the methods developed for deterministic counterparts of the equations. The scaling of the computational cost in the proposed approach is linear in the number of time steps as opposed to the cubic scaling of the direct GP regression solution. We also show how separable covariance functions lead to a finite-dimensional Kalman filtering and RTS smoothing problem, present analytical and numerical examples, and discuss numerical methods for computing the solutions.

1 Introduction

Gaussian process (GP) regression (O'Hagan, 1978; Rasmussen and Williams, 2006) is a Bayesian machine learning paradigm, where the model functions are assumed to be realizations from a Gaussian random process prior. Learning in GP models amounts to computing the posterior process given a set of measurements, and prediction means computing predictive distributions of the function values at new input points. In the usual setting, the GP is constructed by postulating the prior mean function $m_0(\mathbf{x})$ and the prior covariance function $C_0(\mathbf{x}, \mathbf{x}')$ for the Gaussian model functions $f(\mathbf{x})$. Because GP regression is formally equivalent to Gaussian random field based Kriging in geostatistics (Cressie, 1993), all the results presented in this paper are also directly applicable to the corresponding geostatistical models.

GP regression can, in principle, be used for spatio-temporal modeling simply by postulating the mean and covariance functions $m_0(\mathbf{x}, t)$ and $C_0(\mathbf{x}, t; \mathbf{x}', t')$, respectively, for the spatio-temporal process $f(\mathbf{x}, t)$. However, this procedure leads to an unfeasible cubic $\mathcal{O}(M^3\,T^3)$ computational cost, where M is the average number of measurements per time step and T is the number of time points.

Kalman filter (Kalman, 1960; Jazwinski, 1970; Grewal and Andrews, 2001) is a classical algorithm, which can be used for computing the Bayesian solutions to a general class of temporal Gaussian processes observed through a Gaussian linear model. Instead of postulating the mean and covariance functions for the process, the model is constructed as the solution to a linear stochastic differential equation (SDE) such as

$$d\mathbf{f}(t) = \mathbf{A}\,\mathbf{f}(t)\,dt + \mathbf{L}\,d\mathbf{W}(t), \qquad (1)$$

where $\mathbf{f} = (f, df/dt, \ldots, d^{s-1}f/dt^{s-1})$ and $\mathbf{W}(t)$ is a Wiener process. The advantage of this procedure is that the complexity of the approach is linear $\mathcal{O}(T)$ in the number of time steps. Actually, the Kalman filter only provides the forward-time posteriors of the process and to get the full posterior one needs to use the Rauch-Tung-Striebel (RTS) or other type of linear smoother (Rauch et al., 1965; Grewal and Andrews, 2001). Originally, Kalman filter was derived as a com-

Appearing in Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume 22 of JMLR: W&CP 22. Copyright 2012 by the authors.

putationally efficient solution to the Wiener filtering problem (Wiener, 1950), which can be considered as an early version of GP regression.

As the models used in Kalman filtering are also Gaussian processes, one would expect that there would be a connection between GP regression and Kalman filtering. There indeed is, as has recently been explicitly shown by Hartikainen and Särkkä (2010), but only in the case of scalar input. In that case it is possible to interpret the input as time and find a suitable linear SDE such that its covariance function matches that of the GP regression model. The advantage of this Kalman filtering and smoothing formulation is that its computational cost is linear in the number of time steps $\mathcal{O}(T)$ as opposed to the cubic cost $\mathcal{O}(T^3)$ of direct GP regression (Hartikainen and Särkkä, 2010).

Lindgren et al. (2011) recently analyzed the classical link (Whittle, 1954; Matérn, 1960) between stationary stochastic partial differential equations (SPDEs), Gaussian fields, and Gauss-Markov random fields (GMRFs). The methods proposed by Lindgren et al. (2011) can be used for converting between SPDE, covariance function, and GMRF representations of spatio-temporal fields. But the approach does not solve the cubic time scaling problem, because in spatio-temporal case the approach amounts to using the classical conversion procedure for getting the covariance function $C(\mathbf{x}, t)$ and then approximating it with Hilbert space methods. The approach is not linear in time, because the straight-forward application of the classical conversion of SPDEs to covariance functions does not lead to models, which would be stable in forward time (*causal* in signal processing terminology). This means that the models are not Markovian in time series sense and thus usage of Kalman filter and smoother type linear time $\mathcal{O}(T)$ algorithms is not possible. To achieve the linear time complexity $\mathcal{O}(T)$, it is necessary to use a spatio-temporal analog of the spectral factorization approach of Hartikainen and Särkkä (2010). Obviously, when one uses only $\mathcal{O}(T)$ basis functions, the approach of Lindgren et al. (2011) can be made linear in time - but with the cost of very rough approximation.

In this paper, we extend the linear-time temporal GP regression approach (Hartikainen and Särkkä, 2010) to spatio-temporal GP regression models by combining it with the classical conversion procedure (Whittle, 1954; Matérn, 1960; Lindgren et al., 2011). We show how spatio-temporal GP regression can be posed as an infinite-dimensional (or "distributed parameter") Kalman filtering and RTS smoothing problem (Curtain, 1975; Tzafestas, 1978; Omatu and Seinfeld, 1989; Wikle and Cressie, 1999; Cressie and Wikle, 2002), and present a procedure for converting spatio-temporal covariance functions into linear causal evolution type infinite-dimensional stochastic (partial/pseudo) differential equations, where the matrix $\mathbf{A}$ is replaced with a differential or pseudo-differential operator (Lanczos, 1997; Shubin, 1987). We also show how separable covariance functions lead to finite-dimensional models, which can be solved using finite-dimensional Kalman filters and RTS smoothers. We also present simulated and real data examples, and discuss on numerical methods for computing the solutions[1].

The scaling of the proposed approach is linear in the number of time steps $\mathcal{O}(T)$ as opposed to the cubic scaling $\mathcal{O}(T^3)$ of the direct GP regression solution using a general spatio-temporal covariance function. The spatio-temporal complexity of the proposed approach depends on the approximations used and can vary from $\mathcal{O}(M\,T)$ with efficient sparse approximations to $\mathcal{O}(M^3\,T)$ when no sparse approximations are used.

2 From Gaussian Processes to Infinite-Dimensional Filtering and Smoothing

In this section we show how Gaussian process (GP) regression can be seen as infinite-dimensional linear regression, how Kalman filtering and RTS smoothing can be seen as time-varying extensions of linear regression, and finally how Kalman filtering and smoothing can be extended to infinite dimensions. We also discuss on computational methods, which can be used for implementing the methods in practice.

2.1 GP Regression as Infinite-Dimensional Linear Model

Consider the following finite-dimensional linear model

$$\begin{aligned} \mathbf{f} &\sim \mathrm{N}(\mathbf{m}_0, \mathbf{C}_0) \\ \mathbf{y} &= \mathbf{H}\,\mathbf{f} + \mathbf{e}, \end{aligned} \tag{2}$$

where the unknown latent vector is $\mathbf{f} \in \mathbb{R}^s$, $\mathbf{y} \in \mathbb{R}^n$ is the vector of measurements, $\mathbf{H} \in \mathbb{R}^{n \times s}$ is the measurement model matrix (or regressor matrix), and $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \Sigma)$ is a vector of measurement errors. The posterior distribution of $\mathbf{f}$ given $\mathbf{y}$ is now Gaussian $p(\mathbf{f} \,|\, \mathbf{y}) = \mathrm{N}(\mathbf{f} \,|\, \hat{\mathbf{m}}, \hat{\mathbf{C}})$ with the mean and covariance

$$\begin{aligned} \hat{\mathbf{m}} &= \mathbf{m}_0 + \mathbf{C}_0\,\mathbf{H}^T(\mathbf{H}\,\mathbf{C}_0\,\mathbf{H}^T + \Sigma)^{-1}[\mathbf{y} - \mathbf{H}\,\mathbf{m}_0] \\ \hat{\mathbf{C}} &= \mathbf{C}_0 - \mathbf{C}_0\,\mathbf{H}^T(\mathbf{H}\,\mathbf{C}_0\,\mathbf{H}^T + \Sigma)^{-1}\mathbf{H}\,\mathbf{C}_0. \end{aligned} \tag{3}$$

Let's now consider the case, where f is not a vector, but an element of an infinite dimensional Hilbert space $f(\mathbf{x}) \in \mathcal{H}(\mathbb{R}^d)$. The components f_i in the model (2)

[1] For more details, see the supplementary material.

now correspond to values of the function $f(\mathbf{x})$ with different inputs. That is, the index $i \in \{1, \ldots, n\}$ gets replaced with the input value $\mathbf{x} \in \mathbb{R}^d$. With this identification the Gaussian prior $\mathrm{N}(\mathbf{m}_0, \mathbf{C}_0)$ becomes a Gaussian process prior $\mathrm{GP}(m_0(\mathbf{x}), C_0(\mathbf{x}, \mathbf{x}'))$ and the matrix $\mathbf{H}$ becomes a vector of functionals $\boldsymbol{\mathcal{H}} : \mathcal{H}(\mathbb{R}^d) \mapsto \mathbb{R}^n$ (cf. Särkkä, 2011):

$$\begin{aligned} f(\mathbf{x}) &\sim \mathrm{GP}(m_0(\mathbf{x}), C_0(\mathbf{x}, \mathbf{x}')) \\ \mathbf{y} &= \boldsymbol{\mathcal{H}}\, f(\mathbf{x}) + \mathbf{e}, \end{aligned} \tag{4}$$

where we still have $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \Sigma)$. The posterior mean and covariance functions are now given by infinite-dimensional analogs of the Equations (3):

$$\begin{aligned} \hat{m}(\mathbf{x}) &= m_0(\mathbf{x}) + C_0(\mathbf{x}, \mathbf{x}')\boldsymbol{\mathcal{H}}^* \\ &\quad \times \left[\boldsymbol{\mathcal{H}}\, C_0(\mathbf{x}, \mathbf{x}')\, \boldsymbol{\mathcal{H}}^* + \Sigma\right]^{-1} \left[\mathbf{y} - \boldsymbol{\mathcal{H}}\, m_0(\mathbf{x})\right] \\ \hat{C}(\mathbf{x}, \mathbf{x}') &= C_0(\mathbf{x}, \mathbf{x}') - C_0(\mathbf{x}, \mathbf{x}')\, \boldsymbol{\mathcal{H}}^* \\ &\quad \times \left[\boldsymbol{\mathcal{H}}\, C_0(\mathbf{x}, \mathbf{x}')\, \boldsymbol{\mathcal{H}}^* + \Sigma\right]^{-1} \boldsymbol{\mathcal{H}}\, C_0(\mathbf{x}, \mathbf{x}'), \end{aligned} \tag{5}$$

where the operator multiplication from left means operating to variable $\mathbf{x}$ and multiplication from right means operating to $\mathbf{x}'$. The $()^*$ denotes an adjoint, which in practice exchanges the roles of variables $\mathbf{x}$ and $\mathbf{x}'$, and transposes the matrix (cf. Särkkä, 2011).

Note that if we define the measurement model functional $\boldsymbol{\mathcal{H}}\, f(\mathbf{x}) = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$, then the model reduces to

$$\begin{aligned} f(\mathbf{x}) &\sim \mathrm{GP}(m_0(\mathbf{x}), C_0(\mathbf{x}, \mathbf{x}')) \\ y_j &= f(\mathbf{x}_j) + e_j, \qquad j = 1, \ldots, n, \end{aligned} \tag{6}$$

where $(e_1, \ldots, e_n) \sim \mathrm{N}(\mathbf{0}, \Sigma)$, which is just the basic GP regression model (O'Hagan, 1978; Rasmussen and Williams, 2006). It is easy to see that in this case the Equations (5) also reduce to the basic GP regression equations.

In this paper, we concentrate on spatio-temporal processes, which means that the functions are also dependent on another variable t, which can be interpreted as time. Without loss of generality we may assume that the measurements are obtained at certain discrete points of time $t_k, k = 1, \ldots, T$, which do not need to be distinct, and the model can be written in form

$$\begin{aligned} f(\mathbf{x}, t) &\sim \mathrm{GP}(m_0(x, t), C_0(\mathbf{x}, \mathbf{x}'; t, t')) \\ \mathbf{y}_k &= \boldsymbol{\mathcal{H}}_k f(\mathbf{x}, t_k) + \mathbf{e}_k, \end{aligned} \tag{7}$$

where the measurements $\mathbf{y}_k$ have dimension n_k and the measurement noises form an IID sequence $\mathbf{e}_k \sim \mathrm{N}(\mathbf{0}, \Sigma_k)$. The direct GP regression solution to this problem could now be constructed by interpreting time t as an additional input in Equations (5), which would result in an algorithm with cubic computational cost $\mathcal{O}(T^3)$ in number of time steps.

2.2 Infinite-Dimensional Kalman Filtering and Smoothing

Assume that we extend the linear model (2) such that the vector is allowed to change in time according to a linear stochastic differential equation (SDE) model (see, Karatzas and Shreve, 1991; Øksendal, 2003), and a new vector of measurements is obtained at times t_k for $k = 1, \ldots, T$:

$$\begin{aligned} d\mathbf{f}(t) &= \mathbf{A}\, \mathbf{f}(t)\, dt + \mathbf{L}\, d\mathbf{W}(t) \\ \mathbf{y}_k &= \mathbf{H}_k\, \mathbf{f}(t_k) + \mathbf{e}_k, \end{aligned} \tag{8}$$

where $\mathbf{f}(t) \in \mathbb{R}^s$, $\mathbf{A} \in \mathbb{R}^{s \times s}$, $\mathbf{L} \in \mathbb{R}^{s \times q}$, $\mathbf{H}_k \in \mathbb{R}^{n_k \times s}$ are given matrices, $\mathbf{y}_k \in \mathbb{R}^{n_k}$, $\mathbf{e}_k \sim \mathrm{N}(\mathbf{0}, \Sigma_k)$, and $\mathbf{W}(t) \in \mathbb{R}^q$ is a Wiener process with a given diffusion matrix $\mathbf{Q}_c \in \mathbb{R}^{q \times q}$. The prior is assumed to be given as $\mathbf{f}(t_0) \sim \mathrm{N}(\mathbf{m}_0, \mathbf{C}_0)$. As well known, the problem of estimating $\mathbf{f}(t)$ given the measurements can be now solved using the classical Kalman filter and Rauch-Tung-Striebel (RTS) smoother (Kalman, 1960; Rauch et al., 1965; Grewal and Andrews, 2001).

The infinite-dimensional counterpart of the model (8) is the following:

$$\begin{aligned} d\mathbf{f}(\mathbf{x}, t) &= \boldsymbol{\mathcal{A}}\, \mathbf{f}(\mathbf{x}, t)\, dt + \mathbf{L}\, d\mathbf{W}(\mathbf{x}, t) \\ \mathbf{y}_k &= \boldsymbol{\mathcal{H}}_k\, \mathbf{f}(\mathbf{x}, t_k) + \mathbf{e}_k, \end{aligned} \tag{9}$$

where $\mathbf{x} \mapsto f_j(\mathbf{x}, t) \in \mathcal{H}(\mathbb{R}^d)$ for $j = 1, \ldots, s$, $\boldsymbol{\mathcal{A}}$ is a $s \times s$ matrix of linear operators operating on $\mathbf{x}$ with elements $\mathcal{A}_{nm} : \mathcal{H}(\mathbb{R}^d) \mapsto \mathcal{H}(\mathbb{R}^d)$, $\mathbf{L} \in \mathbb{R}^{s \times q}$ is a matrix, $\boldsymbol{\mathcal{H}}_k$ is a $n_k \times s$ matrix of linear functionals operating on $\mathbf{x}$ with elements $\mathcal{H}_{k,nm} : \mathcal{H}(\mathbb{R}^d) \mapsto \mathbb{R}$, $\mathbf{y}_k \in \mathbb{R}^{n_k}$, $\mathbf{e}_k \sim \mathrm{N}(\mathbf{0}, \Sigma_k)$, and $\mathbf{W}(\mathbf{x}, t)$ is a q-dimensional vector of Hilbert space $\mathcal{H}(\mathbb{R}^d)$ valued Wiener processes with the joint diffusion operator $\mathbf{Q}_c(\mathbf{x}, \mathbf{x}')$. The prior is assumed to be given as $\mathbf{f}(\mathbf{x}, t_0) \sim \mathrm{GP}(\mathbf{m}_0(\mathbf{x}), \mathbf{C}_0(\mathbf{x}, \mathbf{x}'))$. The dynamic model in the above equation now is an infinite-dimensional linear Itô stochastic differential equation (Da Prato and Zabczyk, 1992). If the operator $\boldsymbol{\mathcal{A}}$ happens to be a differential operator, the equation becomes an evolution type stochastic partial differential equation (Chow, 2007). However, in this paper we consider a more general class of equations, where the operators are pseudo-differential operators (Shubin, 1987). Note that the GP regression model (4) is obtained as a special case with $\boldsymbol{\mathcal{A}} = \mathbf{0}$, $\mathbf{Q}_c(\mathbf{x}, \mathbf{x}') = \mathbf{0}$, and only one measurement step.

The model can now be converted into an equivalent discrete-time model in analogous manner to the finite-dimensional case (see, e.g., Grewal and Andrews, 2001). First compute the evolution operator $\boldsymbol{\mathcal{U}}(t) = \exp(t\, \boldsymbol{\mathcal{A}})$, where $\exp(\cdot)$ is the operator exponential function. The mild solution to the stochastic equation can now be expressed as (Da Prato and Zabczyk, 1992):

$$\mathbf{f}(\mathbf{x},t) = \boldsymbol{\mathcal{U}}(t-s)\,\mathbf{f}(\mathbf{x},s) + \int_s^t \boldsymbol{\mathcal{U}}(t-\tau)\,\mathbf{L}\,d\mathbf{W}(\mathbf{x},\tau), \tag{10}$$

where t and $s < t$ are arbitrary. The second term above is a GP with the covariance function $\mathbf{Q}(\mathbf{x},\mathbf{x}';t-s) = \int_s^t \boldsymbol{\mathcal{U}}(t-\tau)\,\mathbf{L}\,\mathbf{Q}_c(\mathbf{x},\mathbf{x}')\,\mathbf{L}^T\boldsymbol{\mathcal{U}}^*(t-\tau)\,d\tau$, and thus we can express the model (9) at times t_k as the following discrete-time model (cf. Wikle and Cressie, 1999; Cressie and Wikle, 2002; Gelfand et al., 2010, Part V):

$$\begin{aligned} \mathbf{f}(\mathbf{x},t_k) &= \boldsymbol{\mathcal{U}}(\Delta t_k)\,\mathbf{f}(\mathbf{x},t_{k-1}) + \mathbf{v}_k(\mathbf{x}) \\ \mathbf{y}_k &= \boldsymbol{\mathcal{H}}_k\,\mathbf{f}(\mathbf{x},t_k) + \mathbf{e}_k, \end{aligned} \tag{11}$$

where $\Delta t_k = t_k - t_{k-1}$ and $\mathbf{v}_k(\mathbf{x}) \sim \mathrm{GP}(\mathbf{0},\mathbf{Q}(\mathbf{x},\mathbf{x}';\Delta t_k))$. Note that the discretization above is exact, because it is the mild solution to the infinite-dimensional stochastic differential equation, not an approximation to it. If we want to predict the values at certain new time points t^*, we need to include them as additional measurement-less time points to the discretization above.

The infinite-dimensional Kalman filter and smoother (see, e.g., Tzafestas, 1978; Omatu and Seinfeld, 1989; Cressie and Wikle, 2002) can be now written in the following form, which is formally equivalent to the finite-dimensional case with matrices replaced with operators and functionals:

- *Filtering*: Starting from $\mathbf{m}_0(\mathbf{x})$ and $\mathbf{C}_0(\mathbf{x})$, perform the following for $k = 1,\ldots,T$:

$$\begin{aligned} \mathbf{m}_k^-(\mathbf{x}) &= \boldsymbol{\mathcal{U}}(\Delta t_k)\,\mathbf{m}_{k-1}(\mathbf{x}) \\ \mathbf{C}_k^-(\mathbf{x},\mathbf{x}') &= \boldsymbol{\mathcal{U}}(\Delta t_k)\,\mathbf{C}_{k-1}(\mathbf{x},\mathbf{x}')\,\boldsymbol{\mathcal{U}}^*(\Delta t_k) \\ &\quad + \mathbf{Q}(\mathbf{x},\mathbf{x}';\Delta t_k) \\ \mathbf{m}_k(\mathbf{x}) &= \mathbf{m}_k^-(\mathbf{x}) + \mathbf{C}_k^-(\mathbf{x},\mathbf{x}')\,\boldsymbol{\mathcal{H}}_k^* \\ &\quad \times \left[\boldsymbol{\mathcal{H}}_k\,\mathbf{C}_k^-(\mathbf{x},\mathbf{x}')\,\boldsymbol{\mathcal{H}}_k^* + \Sigma_k\right]^{-1} \\ &\quad \times [\mathbf{y}_k - \boldsymbol{\mathcal{H}}_k\,\mathbf{m}_k^-(\mathbf{x})] \\ \mathbf{C}_k(\mathbf{x},\mathbf{x}') &= \mathbf{C}_k^-(\mathbf{x},\mathbf{x}') - \mathbf{C}_k^-(\mathbf{x},\mathbf{x}')\,\boldsymbol{\mathcal{H}}_k^* \\ &\quad \times \left[\boldsymbol{\mathcal{H}}_k\,\mathbf{C}_k^-(\mathbf{x},\mathbf{x}')\,\boldsymbol{\mathcal{H}}_k^* + \Sigma_k\right]^{-1} \\ &\quad \times \boldsymbol{\mathcal{H}}_k\,\mathbf{C}_k^-(\mathbf{x},\mathbf{x}'). \end{aligned}$$

- *Smoothing*: Starting from $\mathbf{m}_T^s = \mathbf{m}_T$ and $\mathbf{C}_T^s = \mathbf{C}_T$, perform the following for $k = T-1,\ldots,0$:

$$\begin{aligned} \mathbf{G}_k(\mathbf{x}) &= \mathbf{C}_k(\mathbf{x},\mathbf{x}')\,\boldsymbol{\mathcal{U}}^*(\Delta t_k)\,\left[\mathbf{C}_{k+1}^-(\mathbf{x},\mathbf{x}')\right]^{-1} \\ \mathbf{m}_k^s(\mathbf{x}) &= \mathbf{m}_k(\mathbf{x}) + \mathbf{G}_k\,\left[\mathbf{m}_{k+1}^s(\mathbf{x}) - \mathbf{m}_{k+1}^-(\mathbf{x})\right] \\ \mathbf{C}_k^s(\mathbf{x},\mathbf{x}') &= \mathbf{C}_k(\mathbf{x},\mathbf{x}') + \mathbf{G}_k(\mathbf{x}) \\ &\quad \times \left[\mathbf{C}_{k+1}^s(\mathbf{x},\mathbf{x}') - \mathbf{C}_{k+1}^-(\mathbf{x},\mathbf{x}')\right]\,\mathbf{G}_k^*(\mathbf{x}), \end{aligned}$$

where $[]^{-1}$ is interpreted as matrix or operator inverse.

The smoothing pass results in the mean and covariance functions, $\mathbf{m}_k^s(\mathbf{x})$ and $\mathbf{C}_k^s(\mathbf{x},\mathbf{x}')$, which are conditioned to the measurements $\mathbf{y}_1,\ldots,\mathbf{y}_T$. Thus, for example, the marginal posterior of $\mathbf{f}(\mathbf{x}^*,t_k)$, where $\mathbf{x}^*$ is a given test point is

$$\begin{aligned} &p(\mathbf{f}(\mathbf{x}^*,t_k)\,|\,\mathbf{y}_1,\ldots,\mathbf{y}_T) \\ &\quad = \mathrm{N}(\mathbf{f}(\mathbf{x}^*,t_k)\,|\,\mathbf{m}_k^s(\mathbf{x}^*),\mathbf{C}_k^s(\mathbf{x}^*,\mathbf{x}^*)). \end{aligned} \tag{12}$$

With this formulation, the smoothing gives the GP regression predictions only at times t_k, but to get solutions at arbitrary times, it is easy to include additional test time points t^* without measurements to the set of times t_k. The marginal likelihood needed in parameter estimation can be evaluated as $p(\mathbf{y}_1,\ldots,\mathbf{y}_T) = \prod_{k=1}^T \mathrm{N}(\mathbf{y}_k\,|\,\boldsymbol{\mathcal{H}}_k\,\mathbf{m}_k^-(\mathbf{x}),\boldsymbol{\mathcal{H}}_k\,\mathbf{C}_k^-(\mathbf{x},\mathbf{x}')\,\boldsymbol{\mathcal{H}}_k^* + \Sigma_k)$.

2.3 Computational Methods

The actual implementation of the infinite-dimensional Kalman filter and smoother requires computation of the exponential of the operator $\exp(t\,\boldsymbol{\mathcal{A}})$ as well as a few other infinite-dimensional operations. Fortunately, all the involved operations can be performed with the well-known analytical and numerical methods for partial differential equations, evolution equations and pseudo-differential equations (see, e.g., Robinson, 2001; Zinn-Justin, 2002; Pivato, 2010; Lamoureux and Margrave, 2008; Lindgren et al., 2011). Particularly useful methods are basis function expansion based methods such as eigenfunction expansions, Galerkin approximations (e.g. FEM), and point collocation methods. The basis function approach was also used for approximating the space-time Kalman filter by Wikle and Cressie (1999). Methods such as finite-differences (FD) approximations or FFT based spectral methods can also be sometimes used.

One particularly useful special class of models are the models, which are "diagonal" with respect to the input variable $\mathbf{x}$. This means that the operator matrix $\boldsymbol{\mathcal{A}}$ and functional matrices $\boldsymbol{\mathcal{H}}_k$ in the model (9) are in fact constant matrices:

$$\begin{aligned} d\mathbf{f}(\mathbf{x},t) &= \mathbf{A}\,\mathbf{f}(\mathbf{x},t)\,dt + \mathbf{L}\,d\mathbf{W}(\mathbf{x},t) \\ \mathbf{y}_k &= \mathbf{H}_k\,\mathbf{f}(\mathbf{x},t_k) + \mathbf{e}_k. \end{aligned} \tag{13}$$

If we are interested in the values of the process at certain finite set of input points $\mathbf{x}_1,\ldots,\mathbf{x}_n$ only (say, at training and test sets), then we can reduce the above model into a finite-dimensional state space model by defining the augmented state as $\mathbf{f}^a(t) = (\mathbf{f}(\mathbf{x}_1,t),\ldots,\mathbf{f}(\mathbf{x}_n,t))$. The model now becomes finite-dimensional and we can use the conventional finite-dimensional Kalman filtering and smoothing techniques for computation. This approach has been used, for example, in the recent articles of Hartikainen et al. (2011) and Hiltunen et al. (2011).

3 Converting Covariance Functions to Stochastic Equations

In this section, we present a method for converting space-time covariances into equivalent infinite-dimensional stochastic differential equations. We assume that the Gaussian processes are stationary, which means that their covariance functions can be written as functions of single d-dimensional input variable. We use slight abuse of notation and write stationary covariance function $C(\mathbf{x},\mathbf{x}') = C(\mathbf{x}-\mathbf{x}')$ simply as $C(\mathbf{x})$. In the case of spatio-temporal covariances, the stationary covariance functions are denoted as $C(\mathbf{x},t)$.

3.1 General Conversion Procedure

Assume that we have been given a stationary (scalar) covariance function $C(\mathbf{x},t)$ for a spatio-temporal process $f(\mathbf{x},t)$, where $\mathbf{x} \in \mathbb{R}^d$ and $t \in \mathbb{R}$. We now want to form an infinite-dimensional SDE, whose solution (approximately) has the same covariance function. By the Fourier transform we can compute the corresponding spectral density $S(\boldsymbol{\omega}_x,\omega_t)$, where $\boldsymbol{\omega}_x \in \mathbb{R}^d$ and $\omega_t \in \mathbb{R}$. The next task is to find a function $G(i\boldsymbol{\omega}_x, i\omega_t)$, which is rational in variable $i\omega_t$ as follows:

$$G(i\boldsymbol{\omega}_x, i\omega_t) = \frac{b_0(i\boldsymbol{\omega}_x)}{(i\omega_t)^N + a_{N-1}(i\boldsymbol{\omega}_x)\,(i\omega_t)^{N-1} + \cdots + a_0(i\boldsymbol{\omega}_x)}, \tag{14}$$

such that its absolute value approximates the spectral density well: $S(\boldsymbol{\omega}_x,\omega_t) \approx G(i\boldsymbol{\omega}, i\omega_t)\,G(-i\boldsymbol{\omega}, -i\omega_t)$. One practical way to form this kind of approximation – if the spectral density does not already have the suitable form – is to Taylor expand the inverse of spectral density in terms of $(i\omega_t)^2$, which gives $2N$'th order polynomial approximation of the form

$$\frac{1}{S(\boldsymbol{\omega}_x,\omega_t)} \approx c_0(i\boldsymbol{\omega}_x) + c_2(i\boldsymbol{\omega}_x)\,(i\omega_t)^2 + c_4(i\boldsymbol{\omega}_x)\,\omega_t^4 + \cdots \tag{15}$$

Other methods such as orthogonal polynomials or point-wise polynomial fitting could be used equally well. We can then do spectral factorization with respect the variable $i\omega_t$ as was done in Hartikainen and Särkkä (2010).

Once we have obtained the rational approximation (14), we can write down the equation for the (generalized) Fourier transform of $f(\mathbf{x},t)$ formally as follows:

$$F(i\boldsymbol{\omega}_x, i\omega_t) = G(i\boldsymbol{\omega}_x, i\omega_t)\,N(i\boldsymbol{\omega}_x, i\omega_t), \tag{16}$$

where $N(i\boldsymbol{\omega}_x, i\omega_t)$ is the formal Fourier transform of the space-time white noise with unity spectral density. The spectral density of $F(i\boldsymbol{\omega}_x, i\omega_t)$ is now $|F(i\boldsymbol{\omega}_x, i\omega_t)|^2 = G(i\boldsymbol{\omega}_x, i\omega_t)\,G(-i\boldsymbol{\omega}_x, -i\omega_t) \approx S(\boldsymbol{\omega}_x,\omega_t)$ as desired. The inverse Fourier transform $\tilde{f}(i\boldsymbol{\omega}_x, t)$ of $F(i\boldsymbol{\omega}_x, i\omega_t)$ with respect to time can now be implemented by as follows (cf. Hartikainen and Särkkä, 2010):

$$d\tilde{\mathbf{f}}(i\boldsymbol{\omega}_x, t) = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -a_0(i\boldsymbol{\omega}_x) & -a_1(i\boldsymbol{\omega}_x) & \dots & -a_{N-1}(i\boldsymbol{\omega}_x) \end{pmatrix} \times \tilde{\mathbf{f}}(i\boldsymbol{\omega}_x, t)\,dt + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} d\tilde{W}(i\boldsymbol{\omega}_x, t), \tag{17}$$

where the actual process is the first component $\tilde{f} \triangleq \tilde{f}_1$ and $t \mapsto \tilde{W}(i\boldsymbol{\omega}_x, t)$ is a scalar Wiener process with diffusion constant $|b_0(i\boldsymbol{\omega}_x)|^2$.

By taking the inverse Fourier transform $\mathcal{F}_x^{-1}[]$ with respect to the input variable and writing the result in stochastic differential equation form gives the following stochastic evolution equation:

$$d\mathbf{f}(\mathbf{x},t) = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -\mathcal{A}_0 & -\mathcal{A}_1 & \dots & -\mathcal{A}_{N-1} \end{pmatrix} \mathbf{f}(\mathbf{x},t)\,dt + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} dW(\mathbf{x},t), \tag{18}$$

where $W(\mathbf{x},t)$ is a Hilbert space valued Wiener process with stationary diffusion operator $Q_c(\mathbf{x},\mathbf{x}') \triangleq Q_c(\mathbf{x}) = \mathcal{F}_x^{-1}[|b_0(i\boldsymbol{\omega}_x)|^2]$. The operators $\mathcal{A}_j$ are linear operators defined in terms of their Fourier transforms:

$$\begin{aligned} \mathcal{A}_0 &= \mathcal{F}_x^{-1}[a_0(i\boldsymbol{\omega}_x)], \\ \mathcal{A}_1 &= \mathcal{F}_x^{-1}[a_1(i\boldsymbol{\omega}_x)], \\ &\dots \\ \mathcal{A}_{N-1} &= \mathcal{F}_x^{-1}[a_{N-1}(i\boldsymbol{\omega}_x)]. \end{aligned} \tag{19}$$

If the terms $a_j(i\boldsymbol{\omega}_x)$ happen to be rational functions, the operators are integro-differential operators. In particular, if they happen to be polynomials the Equation (18) becomes so called stochastic partial differential equation (SPDE) of the evolution type. If the terms are nor polynomials or rational functions, the operators are so called pseudo-differential operators and the Equation (18) becomes a stochastic pseudo-differential equation or a fractional stochastic equation, which type of equations have been recently studied, for example, by Kelbert et al. (2005); Angulo et al. (2008).

3.2 Separable Covariance Functions

One useful special case is obtained, when the covariance function is separable $C(\mathbf{x},t) = C(\mathbf{x})\, C(t)$. Then the spectral density is also separable $S(\boldsymbol{\omega}_x, \omega_t) = S(\boldsymbol{\omega}_x)\, S(\omega_t)$, and the transfer function (14) can be selected to be of the form

$$G(i\boldsymbol{\omega}_x, i\omega_t) = \frac{b_0(i\boldsymbol{\omega}_x)}{(i\omega_t)^N + a_{N-1}\,(i\omega_t)^{N-1} + \cdots + a_0}, \tag{20}$$

where $|b_0(i\boldsymbol{\omega}_x)|^2 = \text{const} \times S(\boldsymbol{\omega}_x)$, and a_j are constants, which can be easily determined with the one-dimensional procedure presented by Hartikainen and Särkkä (2010). Comparing to Equations (17) and (18) it is now easy to see that we get an equation of the "diagonal" form (13), where the diffusion operator of the Wiener process $W(\mathbf{x},t)$ is constant times $C(\mathbf{x})$. As discussed in Section 2.3, this model can be then reduced to a finite-dimensional state space model and thus the estimation becomes very light.

Example 3.1 (Squared exponential covariance)**.** *Assume that the space-time covariance function is given by the squared exponential (SE) covariance function, which indeed is separable:*

$$\begin{aligned} C(\mathbf{x},t) &= \exp\left(-\alpha_x\,||\mathbf{x}||^2 - \alpha_t\,t^2\right) \\ &= \exp\left(-\alpha_x\,||\mathbf{x}||^2\right)\,\exp\left(-\alpha_t\,t^2\right). \end{aligned} \tag{21}$$

Using the procedure in Section 4.2 of (Hartikainen and Särkkä, 2010) we can now find the state space model parameters $\mathbf{A}$*,* $\mathbf{L}$ *and* q_c *for the time part defined by* $C(t)$ *and* $S(\omega_t)$*. The resulting infinite-dimensional SDE will then be of the form* (13)*, where the diffusion operator of the Wiener process is* $q_c\,\exp\left(-\alpha_x\,||\mathbf{x}||^2\right)$*.*

4 Analytical and Numerical Results

4.1 Cressie & Huang Spatio-Temporal Covariance Function

Consider the stationary covariance function introduced in Example 1 of (Cressie and Huang, 1999):

$$C(\mathbf{x},t) = \frac{\sigma^2}{(a^2t^2+1)^{d/2}}\exp\left(-\frac{b^2||\mathbf{x}||^2}{a^2t^2+1}\right). \tag{22}$$

Taking Fourier transform with respect to $\mathbf{x}$ and t gives the spectral density

$$\begin{aligned} &S(\boldsymbol{\omega}_x,\omega_t) \\ &= \frac{2\sigma^2\pi^{(d+1)/2}}{a\,||\boldsymbol{\omega}_x||\,b^{d-1}}\exp\left(-\frac{||\boldsymbol{\omega}_x||^2}{4b^2}\right)\exp\left(-\frac{b^2}{a^2||\boldsymbol{\omega}_x||^2}\omega_t^2\right). \end{aligned} \tag{23}$$

The fourth order Taylor series expansion approximation can now be formed as follows:

$$\begin{aligned} &\exp\left(\frac{b^2}{a^2||\boldsymbol{\omega}_x||^2}\omega_t^2\right) \approx \frac{1}{2}\left(\frac{b^2}{a^2||\boldsymbol{\omega}_x||^2}\right)^2 \\ &\times\left(2\left(\frac{a^2||\boldsymbol{\omega}_x||^2}{b^2}\right)^2 - 2\left(\frac{a^2||\boldsymbol{\omega}_x||^2}{b^2}\right)(i\omega_t)^2 + (i\omega_t)^4\right). \end{aligned} \tag{24}$$

The roots of the polynomial on the right are given as $r = \pm 2^{1/4}\exp(\pm i\pi/8)\,||\boldsymbol{\omega}_x||\,(a/b)$, and thus the stable roots are $r_s = -2^{1/4}\exp(\pm i\pi/8)\,||\boldsymbol{\omega}_x||\,(a/b)$. Thus, we get the following stable transfer function:

$$G(i\boldsymbol{\omega}_x, i\omega_t) = \frac{b_0(i\boldsymbol{\omega})}{(i\omega_t)^2 + a_1\,(i\omega_t) + a_0}, \tag{25}$$

where

$$\begin{aligned} a_0 &= 2^{1/2}\,||\boldsymbol{\omega}_x||^2\,\left(\frac{a}{b}\right)^2 \\ a_1 &= 2^{5/4}\cos\left(\frac{\pi}{8}\right)\,||\boldsymbol{\omega}_x||\,\left(\frac{a}{b}\right) \\ |b_0(i\boldsymbol{\omega}_x)|^2 &= \left(\frac{4\sigma^2\pi^{(d+1)/2}a^3}{b^{d+5}}\right)||\boldsymbol{\omega}_x||^3\exp\left(-\frac{||\boldsymbol{\omega}_x||^2}{4b^2}\right). \end{aligned} \tag{26}$$

The operators in (18) are now given as

$$\begin{aligned} \mathcal{A}_0 &= \mathcal{F}_x^{-1}[2^{1/2}\,||\boldsymbol{\omega}_x||^2\,(a/b)^2] \\ &= -2^{1/2}\,(a/b)^2\,\nabla^2 \\ \mathcal{A}_1 &= \mathcal{F}_x^{-1}[2^{5/4}\cos(\pi/8)\,||\boldsymbol{\omega}_x||\,(a/b)] \\ &= 2^{5/4}\cos(\pi/8)\,(a/b)\,\sqrt{-\nabla^2}, \end{aligned} \tag{27}$$

where $\nabla^2 = \partial^2/\partial x_1^2 + \cdots + \partial^2/\partial x_d^2$ is the Laplacian, and thus the resulting pseudo-differential evolution equation is of the form

$$\begin{aligned} &d\mathbf{f}(\mathbf{x},t) \\ &= \begin{pmatrix} 0 & 1 \\ 2^{1/2}\,(a/b)^2\nabla^2 & -2^{5/4}\cos(\pi/8)\,(a/b)\,\sqrt{-\nabla^2} \end{pmatrix} \\ &\quad\times\mathbf{f}(\mathbf{x},t)\,dt + \begin{pmatrix}0\\1\end{pmatrix}dW(\mathbf{x},t). \end{aligned} \tag{28}$$

To evaluate the accuracy of the above approximation, we formed a finite dimensional approximation to the

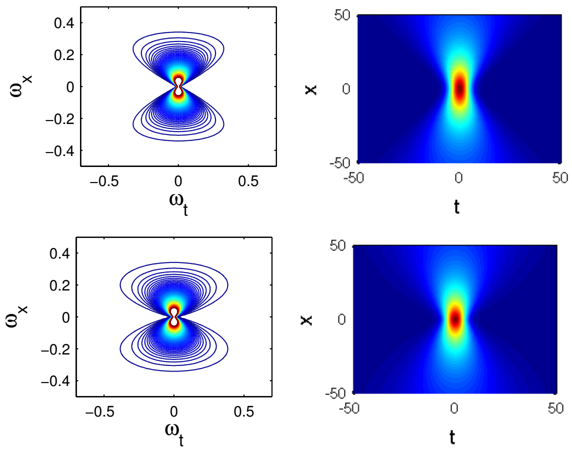

Figure 1: Spectrum of Cressie & Huang model with $d = 1$, $a = 1/10$, $b = 1/10$, $\sigma = 10$ (top left) and the covariance function (top right). Spectrum of the stochastic equation (bottom from left), and its covariance function (bottom right).

equation on finite range $x \in [-100, 100]$ by projecting the process $f(x, t)$ onto the first 50 eigenfunctions of the Laplace operator with Dirichlet boundary conditions.

The original spectral density and covariance function together with the spectral density and covariance function of the above equation with scalar x are shown in Figure 1. It can be seen that the spectra and covariance functions indeed are very similar, although, slight approximation induced differences can be seen. The predictions of the GP regression solution and its Kalman/RTS approximation are shown in Figure 2 and the predictions are indeed very similar. With M eigenfunctions and T measurements, the computations need by the Kalman/RTS approach are of the order $M^3\,T$, whereas the requirements for GP solution are of the order T^3. Thus with the used values $M = 50$ and $T = 500$ then Kalman/RTS is a couple of orders of magnitude lighter than the GP regression solution.

4.2 Modeling of US Monthly Precipitation and Temperature Data

As a real world modeling problem we consider spatio-temporal regression of monthly precipitation and temperature minimum/maximum data[2] collected in the US from years 1895-1997. There are 11918 measurements stations for the precipitation data and 8125 for the temperatures. Subsets of this data were used by Paciorek and Schervish (2006); Vanhatalo and Vehtari (2008) to assess spatial regression models. High fraction of the measurements is missing, and our aim is to fill out the missing measurements by taking account of the spatio-temporal correlations in the data. As the size of original data is very large we focus on (roughly) the same subset of data as used by Paciorek and Schervish (2006). The subset is collected from a rectangular area ($[-109.5, -101] \times [36.5, 41.5]$ lon/lat) around Colorado and comprises of 502 stations for the precipitation and 423 for the temperature readings. The total number of measurements in the subset are 372873 for precipitation, 336156 for maximum temperature and 336720 for minimum temperature.

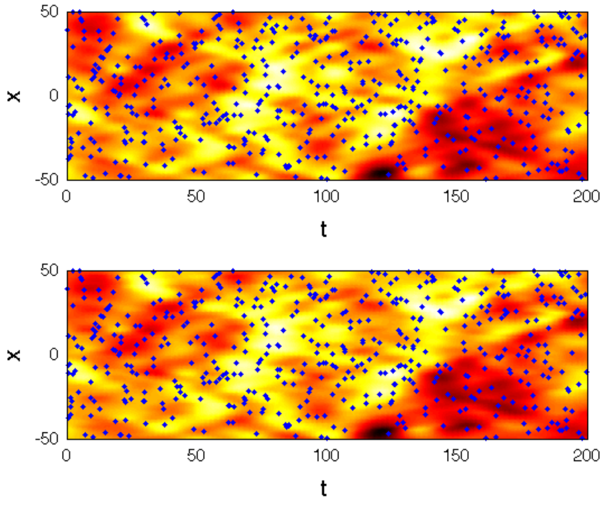

Figure 2: Prediction with Cressie & Huang covariance function with GP regression formulas (left), and with eigenfunction expansion approximation to the infinite-dimensional Kalman filter and RTS smoother (right). The 500 training points are shown with blue dots.

We used 10-fold cross-validation for comparing the predictive performance of models with and without a temporal model. The baseline is independent GP (IGP) for each month separately. For the model that takes into account the temporal dynamics we implemented a spatio-temporal GP (STGP) with a separable covariance function $C(\mathbf{x}, t) = C_x(\mathbf{x})\,C_t(t)$, where the spatial covariance function $C_x(\mathbf{x})$ was a Matérn covariance with smoothness parameter $\nu_x = 3/2$. For the temporal covariance $C_t(t)$ we also used the Matérn class, and tested two different smoothness parameters $\nu_t \in \{1/3, 3/2\}$. To speed up the computations we implemented sparse approximations for both types of models. In particular, we focused on the *fully independent conditional* (FIC) approximation (see, e.g., Quiñonero-Candela and Rasmussen, 2005), which was recently considered for separable spatio-temporal GPs by Hartikainen et al. (2011). The model here is the same, with the exception that Hartikainen et al. (2011) considered non-Gaussian likelihoods whereas we can

[2] http://www.image.ucar.edu/GSP/Data/US.monthly.met/

now assume a Gaussian noise model for the precipitation and temperature measurements. The hyperparameters of both models were optimized with respect to marginal likelihood.

Table 1: Cross-validation based approximations of MSEs for different models of the US monthly precipitation and temperature minima/maxima. The errors are with respect to normalized measurements.

GP	Prec.	Tmax.	Tmin.
STGP(1/2)	**0.22 (0.01)**	**0.029 (0.003)**	**0.028 (0.003)**
STGP(3/2)	0.25 (0.02)	0.034 (0.003)	0.032 (0.004)
IGP	0.30 (0.02)	0.065 (0.003)	0.050 (0.006)

Table 1 shows the estimated predictive mean squared error (MSE) values and their standard errors for all the considered models with 256 inducing points. The spatio-temporal GP with temporal smoothness parameter $\nu = 1/2$ clearly is the best model in all the cases. Johns et al. (2003) discussed that utilization of temporal information might reduce MSE by $1\% - 2\%$ for the precipitation data. Based on our results, in the considered subset, the temporal information reduces error at least 20% in the precipitation and even more in the temperatures. Figure 3 shows the MSE values of all the considered models as a function of number of inducing points m, which were placed in a regular grid over the data. It can be seen that although the number of inducing points indeed affects the MSE values, taking the temporal information into account always reduces the error.

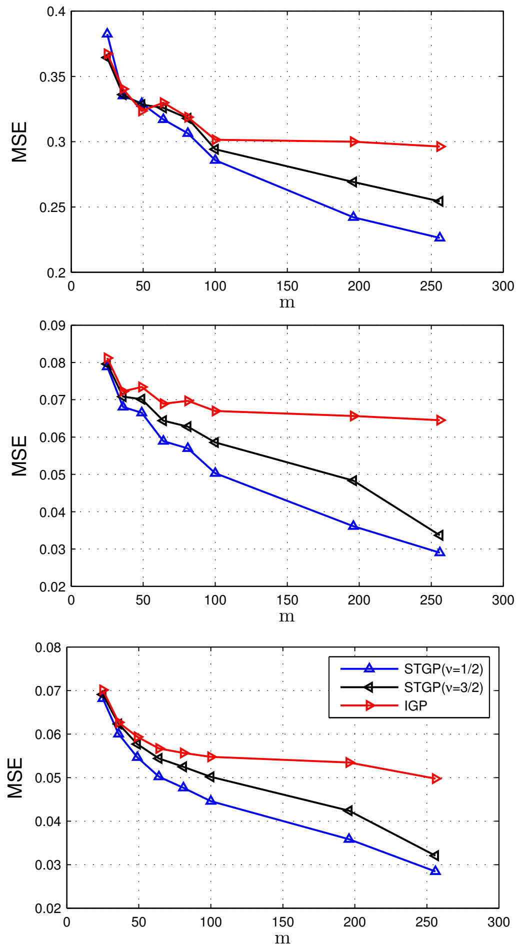

Figure 3: Prediction MSEs as function of number of inducing points for precipitation (top), maximum temperature (middle) and minimum temperature (bottom) with the US monthly data.

5 Conclusion and Discussion

In this paper, we have shown how spatio-temporal Gaussian process (GP) regression can be formulated as an infinite-dimensional Kalman filtering and RTS smoothing problem, and presented a method for converting spatio-temporal covariance functions to infinite-dimensional stochastic differential equations. Using simulated and real-world data we have shown that the proposed method is useful also in practice. The clear advantage of the method is that the computational cost is linear with respect to the number of time steps in contrast to the cubic scaling of the direct GP regression. The disadvantage is that the resulting stochastic equations can be quite complicated and we often need to resort to approximations such as basis function expansions. To further speed up the computations it is also possible to combine the proposed approach with sparse approximations as was done in the US monthly precipitation and temperature data example presented in this paper. Although in this paper we have used the terminology of GP regression, the results also apply to the corresponding Gaussian random field based enviromental and Kriging models used in geostatistics (Cressie, 1993; Lindgren et al., 2011), because the mathematical formulation of the models is the same.

Acknowledgments

The authors acknowledge the financial support from Academy of Finland's Centre of Excellence in Computational Complex Systems Research, Finnish Doctoral Programme in Computational Sciences (FICS), Finnish Foundation for Technology Promotion (TES), and Emil Aaltonen Foundation.

References

Angulo, J. M., Kelbert, M. Y., Leonenko, N. N., and Ruiz-Medina, M. D. (2008). Spatiotemporal random fields associated with stochastic fractional

Helmholtz and heat equations. *Stoch Environ Res Risk Assess*, 22:S3–S13.

Chow, P.-L. (2007). *Stochastic Partial Differential Equations*. Chapman & Hall/CRC.

Cressie, N. and Huang, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *JASA*, 94(448):1330–1340.

Cressie, N. and Wikle, C. K. (2002). Space-time Kalman filter. In El-Shaarawi, A. H. and Piegorsch, W. W., editors, *Encyclopedia of Environmetrics*, volume 4, pages 2045–2049. John Wiley & Sons, Ltd, Chichester.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley.

Curtain, R. (1975). A survey of infinite-dimensional filtering. *SIAM Review*, 17(3):395–411.

Da Prato, G. and Zabczyk, J. (1992). *Stochastic Equations in Infinite Dimensions*. Cambridge University Press.

Gelfand, A. E., Diggle, P. J., Montserrat, and Guttorp, P. (2010). *Handbook of Spatial Statistics*. CRC Press.

Grewal, M. S. and Andrews, A. P. (2001). *Kalman Filtering, Theory and Practice Using MATLAB*. Wiley Interscience.

Hartikainen, J., Riihimäki, J., and Särkkä, S. (2011). Sparse spatio-temporal Gaussian processes with general likelihoods. In *Proceedings of ICANN'11*.

Hartikainen, J. and Särkkä, S. (2010). Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *Proceedings of MLSP*.

Hiltunen, P., Särkkä, S., Nissila, I., Lajunen, A., and Lampinen, J. (2011). State space regularization in the nonstationary inverse problem for diffuse optical tomography. *Inverse Problems*, 27:025009.

Jazwinski, A. (1970). *Stochastic Processes and Filtering Theory*. Academic Press.

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

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME, Journal of Basic Engineering*, 82:35–45.

Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*. Springer.

Kelbert, M., Leonenko, N., and Ruiz-Medina, M. D. (2005). Fractional random fields associated with stochastic fractional heat equations. *Adv Appl Prob*, 37:108–133.

Lamoureux, M. and Margrave, G. (2008). An introduction to numerical methods of pseudodifferential operators. In Feichtinger, H. G., Rodino, L., and Wong, M. W., editors, *Pseudo-differential operators : quantization and signals*, pages 79–133. Springer.

Lanczos, C. (1997). *Linear Differential Operators*. Dover.

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. *JRSS B*, 73(4):423–498.

Matérn, B. (1960). Spatial variation. Technical report, Meddelanden från Statens Skogforskningsinstitut. Band 49 - Nr 5.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction (with discussion). *JRSS B*, 40(1):1–42.

Øksendal, B. (2003). *Stochastic Differential Equations: An Introduction with Applications*. Springer, 6th edition.

Omatu, S. and Seinfeld, J. H. (1989). *Distributed Parameter Systems: Theory and Applications*. Clarendon Press / Ohmsha.

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

Pivato, M. (2010). *Linear Partial Differential Equations and Fourier Theory*. Cambridge University Press.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *JMLR*, 6:1939–1959.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Rauch, H. E., Tung, F., and Striebel, C. T. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450.

Robinson, J. C. (2001). *Infinite-Dimensional Dynamic Systems*. Cambridge University Press.

Särkkä, S. (2011). Linear operators and stochastic partial differential equations in Gaussian process regression. In *Proceedings of ICANN'11*.

Shubin, M. A. (1987). *Pseudodifferential operators and spectral theory*. Springer-Verlag.

Tzafestas, S. G. (1978). Distributed parameter state estimation. In Ray, W. H. and Lainiotis, D. G., editors, *Distributed Parameter Systems*. Dekker.

Vanhatalo, J. and Vehtari, A. (2008). Modelling local and global phenomena with sparse Gaussian processes. In *Proceedings of UAI*.

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

Wiener, N. (1950). *Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*. John Wiley & Sons.

Wikle, C. K. and Cressie, N. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, 86(4):815–829.

Zinn-Justin, J. (2002). *Quantum Field Theory and Critical Phenomena*. Oxford, 4th edition.

SUPPLEMENTARY MATERIAL FOR “Infinite-Dimensional Kalman Filtering Approach to Spatio-Temporal Gaussian Process Regression”

1 Introduction

1.1 Wiener Process and White Noise

In the actual paper, we have denoted stochastic differential equations in Itô notation (cf. Karatzas and Shreve, 1991; Øksendal, 2003) such as

$$d\mathbf{f}(t) = \mathbf{A}\,\mathbf{f}(t)\,dt + \mathbf{L}\,d\mathbf{W}(t), \tag{1}$$

where $\mathbf{W}(t)$ is a Wiener process (or Brownian motion) with diffusion matrix $\mathbf{Q}_c$. The Wiener process is a Gaussian process with statistics:

$$\begin{aligned} \mathrm{E}[\mathbf{W}(t)] &= 0 \\ \mathrm{E}[\mathbf{W}(t)\,\mathbf{W}^T(s)] &= \mathbf{Q}_c\,\min(s,t)\,. \end{aligned} \tag{2}$$

In this supplementary material we will rewrite the equation (1) in differential equation form:

$$d\mathbf{f}(t)/dt = \mathbf{A}\,\mathbf{f}(t) + \mathbf{L}\,\mathbf{w}(t), \tag{3}$$

where the driving process $\mathbf{w}(t)$ is a Gaussian white noise with statistics

$$\begin{aligned} \mathrm{E}[\mathbf{w}(t)] &= 0 \\ \mathrm{E}[\mathbf{w}(t)\,\mathbf{w}^T(s)] &= \mathbf{Q}_c\,\delta(s-t), \end{aligned} \tag{4}$$

and can be considered as the formal derivative of Wiener process $\mathbf{w}(t) = d\mathbf{W}(t)/dt$. Here $\mathbf{Q}_c$ is called the spectral density of the white noise process. The space-time white noise can be defined in analogous manner.

The white noise notation is very convenient in practical computations, because in many cases the differential equations can be treated as if they were deterministic differential equations. For this reason this notation is often preferred in engineering literature (cf. Jazwinski, 1970; Grewal and Andrews, 2001). However, it is important to make sure that every operation is indeed valid in rigorous Itô calculus sense (Karatzas and Shreve, 1991; Øksendal, 2003), and treat the white noise notation only as a convenient notation for the actual Itô calculus in operation. To emphasis the actual meaning of the equations, we have chosen to use the Itô notation in the paper itself.

The background of the notation is that in rigorous sense, we cannot directly define differential equations driven by a white noise such as (3). Let's formally integrate the equation (3), which gives an integral equation of the form

$$\mathbf{f}(t) - \mathbf{f}(t_0) = \int_{t_0}^{t} \mathbf{A}\,\mathbf{f}(t)\,dt + \int_{t_0}^{t} \mathbf{L}\,\mathbf{w}(t)\,dt. \tag{5}$$

Now the last integral cannot be defined as Riemann integral, because the white noise process is formally non-continuous everywhere. However, it can be defined as so called Itô stochastic integral (see, e.g. Karatzas and Shreve, 1991; Øksendal, 2003) provided that we interpret the term $\mathbf{w}(t)\,dt$ as increment of Wiener process $\mathbf{W}(t)$. In Itô formalism the equation can be written in form

$$\mathbf{f}(t) - \mathbf{f}(t_0) = \int_{t_0}^{t} \mathbf{A}\,\mathbf{f}(t)\,dt + \int_{t_0}^{t} \mathbf{L}\,d\mathbf{W}, \tag{6}$$

where $d\mathbf{W}$ is the Wiener process increment. The second integral is now stochastic integral with respect to the stochastic "measure" $\mathbf{W}(t)$, the Wiener process. If we drop the integral signs and consider small values of $t - t_0$, the equation can be written in the more compact form (1), which is the most common notation for Itô stochastic differential equations in stochastics literature. The solution $\mathbf{f}(t)$ of an Itô stochastic differential equation is called an Itô process. Note that the equation can be formally written as

$$d\mathbf{f}(t)/dt = \mathbf{A}\,\mathbf{f}(t) + \mathbf{L}\,d\mathbf{W}/dt, \tag{7}$$

and comparing to Equation (3) reveals that the white noise process can be considered as the formal derivative of Wiener process $d\mathbf{W}/dt$. However, a slightly problematic thing is that the Wiener process is everywhere non-differentiable, and this causes appearance of the delta function in the covariance of white noise.

For the above reasons we also use the Itô notation for infinite-dimensional stochastic differential equations in the actual paper, because there the situation is analogous to the finite-dimensional case. In this supplement we use the white noise notation, because it is easier when doing the actual analytic calculations.

1.2 Multi-Dimensional Fourier Transform

The Fourier transform of function $f(\mathbf{x}) \;:\; \mathbb{R}^d \mapsto \mathbb{R}$ is here defined as

$$\mathcal{F}[f](\boldsymbol{\omega}) = \int_{\mathbb{R}^d} f(\mathbf{x})\,\exp(-\mathrm{i}\,\boldsymbol{\omega}^T\,\mathbf{x})\,d\mathbf{x}. \tag{8}$$

The inverse transform is

$$\mathcal{F}^{-1}[F](\mathbf{x}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} F(\boldsymbol{\omega})\,\exp(\mathrm{i}\,\boldsymbol{\omega}^T\,\mathbf{x})\,d\boldsymbol{\omega}. \tag{9}$$

where $F(\boldsymbol{\omega}) = \mathcal{F}[f](\boldsymbol{\omega})$. Fourier transforms are rarely explicitly computed, but precomputed tables are often used instead (see, e.g. Råde and Westergren, 2004). One-dimensional Fourier transform pairs have been extensively tabulated in literature and because $\exp(\pm \mathrm{i}\, \boldsymbol{\omega}^T \mathbf{x}) = \prod_j \exp(\pm \mathrm{i}\, \omega_j \, x_j)$ multi-dimensional Fourier transforms can be computed as sequential application of single-dimensional transforms. The Fourier transform of a vector valued function can be computed by applying Fourier transform to each of the components of the vector separately.

The important properties, which make Fourier transform particularly useful for solving linear ordinary and partial differential equations are the following:

- Linearity: If $f(\mathbf{x})$ and $g(\mathbf{x})$ are arbitrary functions and $a, b \in \mathbb{R}$ are constants, then:
$$\mathcal{F}[a\, f + b\, g] = a\, \mathcal{F}[f] + b\, \mathcal{F}[g]. \tag{10}$$

- Derivative: If $f(\mathbf{x})$ is a k times differentiable function, defined on whole space $\mathbb{R}^d$ and vanishing at infinity, then the Fourier transform of the partial derivative $\partial^k f / \partial x_i^k$ is
$$\mathcal{F}[\partial^k f / \partial x_i^k] = (\mathrm{i}\, \omega_i)^k \, \mathcal{F}[f]. \tag{11}$$
That is, the Fourier transform maps derivatives to polynomials and thus transforms ordinary and partial differential equations into algebraic equations.

- Convolution: The convolution of functions $f(\mathbf{x})$ and $g(\mathbf{x})$ defined on whole space $\mathbb{R}^d$ as above can be defined as
$$(f * g)(\mathbf{x}) = \int_{R^d} f(\mathbf{x} - \mathbf{x}')\, g(\mathbf{x}')\, d\mathbf{x}'. \tag{12}$$
The Fourier transform of the convolution is then the product of Fourier transforms of f and g:
$$\mathcal{F}[f * g] = \mathcal{F}[f]\, \mathcal{F}[g]. \tag{13}$$

The Fourier transform is also useful in computing the covariance functions of stochastic ordinary and partial differential equations due to the following properties:

- Wiener-Khinchin: If $f(\mathbf{x})$ is a zero mean wide sense stationary random field with covariance function
$$C_f(\mathbf{u}) = \mathrm{E}[f(\mathbf{x})\, f(\mathbf{x} + \mathbf{u})], \tag{14}$$
then the spectral density $S_f(\boldsymbol{\omega})$ of the process $f(\mathbf{x})$ is the Fourier transform of $C_f(\mathbf{u})$:
$$S_f(\boldsymbol{\omega}) = \mathcal{F}[C_f]. \tag{15}$$

- If $h(\mathbf{x})$ is a function and $H(\mathrm{i}\boldsymbol{\omega})$ is Fourier transform (i.e., the transfer function), then the spectral density of the convolution process $g(\mathbf{x}) = h(\mathbf{x}) * f(\mathbf{x})$ is

$$S_g(\boldsymbol{\omega}) = H(\mathrm{i}\boldsymbol{\omega})\, S_f(\boldsymbol{\omega})\, H(-\mathrm{i}\boldsymbol{\omega}) = |H(\mathrm{i}\boldsymbol{\omega})|^2\, S_f(\boldsymbol{\omega}). \tag{16}$$

The Gaussian spatial white noise process can be defined as a random field $w(\mathbf{x})$ with the properties:

$$\begin{aligned} \mathrm{E}[w(\mathbf{x})] &= 0 \\ \mathrm{E}[w(\mathbf{x})\, w(\mathbf{x}+\mathbf{u})] &= q\, \delta(\mathbf{u}). \end{aligned} \tag{17}$$

The spectral density of the white noise process can be obtained as the Fourier transform of the covariance function $C_w(\mathbf{u}) = q\,\delta(\mathbf{u})$ and it is given as

$$S_w(\boldsymbol{\omega}) = q. \tag{18}$$

Due to this property the parameter q or its matrix equivalent in the definition of white noise is often called the spectral density of the white noise process.

In this document and in the paper write we stationary covariance function $C(\mathbf{x}, \mathbf{x}') = C(\mathbf{x} - \mathbf{x}')$ simply as $C(\mathbf{x})$. In the case of spatio-temporal covariances, the stationary covariance functions are denoted as $C(\mathbf{x}, t)$.

2 Details of Squared Exponential Covariance Function Example

The squared exponential (or exponential of square) class of covariance functions has the form

$$C(\mathbf{x}) = \exp\left(-\alpha\, \mathbf{x}^2\right), \tag{19}$$

where in the parameterization of Rasmussen and Williams (2006) we have $\alpha = 1/(2L^2)$. If we rename one of the input as t, and use separate scales for time and input, we get

$$\begin{aligned} C(\mathbf{x}, t) &= \exp\left(-\alpha_x\, \mathbf{x}^2 - \alpha_t\, t^2\right) \\ &= \exp\left(-\alpha_x\, \mathbf{x}^2\right)\, \exp\left(-\alpha_t\, t^2\right) \end{aligned} \tag{20}$$

which can be seen to be separable in space and time. The corresponding spectral density is also separable

$$S(\boldsymbol{\omega}_x, \omega_t) = \left(\frac{\pi}{\alpha_x}\right)^{d/2} \exp\left(-\frac{\boldsymbol{\omega}_x^2}{4\alpha_x}\right) \left(\frac{\pi}{\alpha_t}\right)^{1/2} \exp\left(-\frac{\omega_t^2}{4\alpha_t}\right) \tag{21}$$

Following the procedure presented by Hartikainen and Särkkä (2010) we can now approximate the last term with a polynomial in ω_t^2:

$$\exp\left(-\frac{\omega_t^2}{4\alpha_t}\right) \approx \frac{1}{a_0 + a_1\, (i\omega_t)^2 + \cdots + a_N\, (i\omega)^{2N}}. \tag{22}$$

We can then form the spectral factorization, which will gives

$$\begin{aligned}
&\frac{1}{a_0 + a_1\,(i\omega_t)^2 + \cdots + a_N\,(i\omega)^{2N}} \\
&= \underbrace{\left(\frac{1}{b_0 + b_1\,(i\omega_t) + \cdots + b_N\,(i\omega_t)^N}\right)}_{H_t(i\omega_t)} \underbrace{\left(\frac{1}{b_0 + b_1\,(-i\omega_t) + \cdots + b_N\,(-i\omega_t)^N}\right)}_{H_t(-i\omega_t)}
\end{aligned} \tag{23}$$

where $H_t(i\omega_t)$ has poles only in the upper half plane. Thus we get the approximation

$$S(\boldsymbol{\omega}_x, \omega_t) \approx \hat{S}(\boldsymbol{\omega}_x, \omega_t) = |H_t(i\omega_t)|^2\, S_x(\boldsymbol{\omega}_x), \tag{24}$$

where

$$S_x(\boldsymbol{\omega}_x) = \left(\frac{\pi}{\alpha_x}\right)^{d/2} \left(\frac{\pi}{\alpha_t}\right)^{1/2} \exp\left(-\frac{\boldsymbol{\omega}_x^2}{4\alpha_x}\right). \tag{25}$$

Let $\boldsymbol{\omega}_x$ be fixed and consider the process $\tilde{f}$ satisfying the stochastic differential equation

$$b_0\,\tilde{f}(\boldsymbol{\omega}_x, t) + b_1\,\frac{\partial \tilde{f}(\boldsymbol{\omega}_x, t)}{\partial t} + \cdots + b_N\,\frac{\partial^N \tilde{f}(\boldsymbol{\omega}_x, t)}{\partial t^N} = \tilde{w}(\boldsymbol{\omega}_x, t), \tag{26}$$

where $t \mapsto \tilde{w}(\boldsymbol{\omega}_x, t)$ is a white noise process with spectral density $S_x(\boldsymbol{\omega}_x)$. The process now has the spectral density, which was defined in the Equation (24). Taking inverse Fourier transform with respect to the space then implies that the process satisfying the stochastic equation

$$b_0\,f(\mathbf{x}, t) + b_1\,\frac{\partial f(\mathbf{x}, t)}{\partial t} + \cdots + b_N\,\frac{\partial^N f(\mathbf{x}, t)}{\partial t^N} = w(\mathbf{x}, t), \tag{27}$$

where $w(\mathbf{x}, t)$ is a time-white process with spatial spectral density (25), and thus exponential covariance function, has the spectral density (24) and thus approximately the covariance function (20).

If we define $\mathbf{f} = (f, \partial f/\partial t, \ldots, \partial^{N-1} f/\partial t^{N-1})$, it is easy to see that the above equation can be written in form

$$\frac{\partial \mathbf{f}(\mathbf{x}, t)}{\partial t} = \mathbf{A}\,\mathbf{f}(\mathbf{x}, t) + \mathbf{L}\,w(\mathbf{x}, t) \tag{28}$$

where $\mathbf{A}$ and $\mathbf{L}$ are constant matrices.

3 Details of the Cressie & Huang Example

Consider the stationary covariance function introduced in Example 1 of Cressie and Huang (1999):

$$C(\mathbf{x}, t) = \frac{\sigma^2}{(a^2 t^2 + 1)^{d/2}} \exp\left(-\frac{b^2 ||\mathbf{x}||^2}{a^2 t^2 + 1}\right). \tag{29}$$

The spectral density is Gaussian in space and thus we get the spatial Fourier transform easily:

$$\begin{aligned} \mathcal{F}_x[C(\mathbf{x},t)] &= \frac{\sigma^2 \pi^{d/2}}{b^d} \exp\left(-\frac{a^2 t^2 + 1}{4b^2} ||\boldsymbol{\omega}_x||^2 \right) \\ &= \frac{\sigma^2 \pi^{d/2}}{b^d} \exp\left(-\frac{||\boldsymbol{\omega}_x||^2}{4b^2} \right) \exp\left(-\frac{a^2 ||\boldsymbol{\omega}_x||^2}{4b^2} t^2 \right). \end{aligned} \tag{30}$$

Taking Fourier transform with respect to t is again a Gaussian transform for the last term, which gives the spectral density

$$\begin{aligned} S(\boldsymbol{\omega}_x, \omega_t) &= \frac{\sigma^2 \pi^{d/2}}{b^d} \exp\left(-\frac{||\boldsymbol{\omega}_x||^2}{4b^2} \right) \left(\frac{2b\,\pi^{1/2}}{a\,||\boldsymbol{\omega}_x||} \right) \exp\left(-\frac{b^2}{a^2 ||\boldsymbol{\omega}_x||^2} \omega_t^2 \right) \\ &= \frac{2\sigma^2 \pi^{(d+1)/2}}{a\,||\boldsymbol{\omega}_x||\, b^{d-1}} \exp\left(-\frac{||\boldsymbol{\omega}_x||^2}{4b^2} \right) \exp\left(-\frac{b^2}{a^2 ||\boldsymbol{\omega}_x||^2} \omega_t^2 \right). \end{aligned} \tag{31}$$

Let's form the following Taylor series approximation to the inverse of the last term, write it in terms of $i\omega_t$ and factor out the highest order term:

$$\begin{aligned} &\exp\left(\frac{b^2}{a^2 ||\boldsymbol{\omega}_x||^2} \omega_t^2 \right) \\ &\approx 1 + \left(\frac{b^2}{a^2 ||\boldsymbol{\omega}_x||^2} \right) \omega_t^2 + \frac{1}{2} \left(\frac{b^2}{a^2 ||\boldsymbol{\omega}_x||^2} \right)^2 \omega_t^4 \\ &= 1 - \left(\frac{b^2}{a^2 ||\boldsymbol{\omega}_x||^2} \right) (i\omega_t)^2 + \frac{1}{2} \left(\frac{b^2}{a^2 ||\boldsymbol{\omega}_x||^2} \right)^2 (i\omega_t)^4 \\ &= \frac{1}{2} \left(\frac{b^2}{a^2 ||\boldsymbol{\omega}_x||^2} \right)^2 \left(2 \left(\frac{a^2 ||\boldsymbol{\omega}_x||^2}{b^2} \right)^2 - 2 \left(\frac{a^2 ||\boldsymbol{\omega}_x||^2}{b^2} \right) (i\omega_t)^2 + (i\omega_t)^4 \right) \end{aligned} \tag{32}$$

The roots of the polynomial on the right are given as

$$r = \pm 2^{1/4} \exp(\pm i\pi/8)\, ||\boldsymbol{\omega}_x||\, (a/b), \tag{33}$$

and thus the stable roots are

$$r_s = -2^{1/4} \exp(\pm i\pi/8)\, ||\boldsymbol{\omega}_x||\, (a/b). \tag{34}$$

By expanding the corresponding polynomial, we get the following:

$$(i\omega_t)^2 + 2^{5/4} \cos(\pi/8)\, ||\boldsymbol{\omega}_x||\, (a/b)\, (i\omega_t) + 2^{1/2}\, ||\boldsymbol{\omega}_x||^2\, (a/b)^2. \tag{35}$$

Thus, if we define

$$H(i\boldsymbol{\omega}_x, i\omega_t) = \frac{1}{(i\omega_t)^2 + 2^{5/4} \cos(\pi/8)\, ||\boldsymbol{\omega}_x||\, (a/b)\, (i\omega_t) + 2^{1/2}\, ||\boldsymbol{\omega}_x||^2\, (a/b)^2}. \tag{36}$$

then H is a time-stable transfer function such that

$$S(\boldsymbol{\omega}_x, \omega_t) \approx H(i\boldsymbol{\omega}_x, i\omega_t)\, S_w(\boldsymbol{\omega}_x)\, H(-i\boldsymbol{\omega}_x, -i\omega_t) \tag{37}$$

where

$$\begin{aligned} S_w(\boldsymbol{\omega}_x) &= \frac{2\sigma^2 \pi^{(d+1)/2}}{a\,||\boldsymbol{\omega}_x||\, b^{d-1}} \exp\left(-\frac{||\boldsymbol{\omega}_x||^2}{4b^2}\right) 2 \left(\frac{a^2 ||\boldsymbol{\omega}_x||^2}{b^2}\right)^2 \\ &= \left(\frac{4\sigma^2 \pi^{(d+1)/2} a^3}{b^{d+5}}\right) ||\boldsymbol{\omega}_x||^3 \exp\left(-\frac{||\boldsymbol{\omega}_x||^2}{4b^2}\right) \end{aligned} \tag{38}$$

Now let $w(\mathbf{x}, t)$ be a time-white Gaussian process with spectral density function $Q_w(\mathbf{x}) = \mathcal{F}_x^{-1}[S_w(\boldsymbol{\omega}_x)]$ and define the operators

$$\begin{aligned} \mathcal{A}_0 &= \mathcal{F}_x^{-1}[2^{1/2}\, ||\boldsymbol{\omega}_x||^2\, (a/b)^2] \\ \mathcal{A}_1 &= \mathcal{F}_x^{-1}[2^{5/4} \cos(\pi/8)\, ||\boldsymbol{\omega}_x||\, (a/b)], \end{aligned} \tag{39}$$

then the process $f(\mathbf{x}, t)$ approximately has the covariance function $C(\mathbf{x}, t)$:

$$\frac{\partial^2 f(\mathbf{x}, t)}{\partial t^2} + \mathcal{A}_1 \frac{\partial f(\mathbf{x}, t)}{\partial t} + \mathcal{A}_0 f(\mathbf{x}, t) = w(\mathbf{x}, t). \tag{40}$$

The first of the operators is just

$$\mathcal{A}_0 = 2^{1/2}\ (a/b)^2\, \mathcal{F}_x^{-1}[||\boldsymbol{\omega}_x||^2] = -2^{1/2}\ (a/b)^2\, \nabla^2 \tag{41}$$

The second operator can be written as

$$\mathcal{A}_1 = 2^{5/4} \cos(\pi/8)\, (a/b)\, \mathcal{F}_x^{-1}[||\boldsymbol{\omega}_x||] = 2^{5/4} \cos(\pi/8)\, (a/b)\, \sqrt{-\nabla^2} \tag{42}$$

In numerical computations the operator square root can be usually easily implemented. Thus the resulting pseudo-differential evolution equation is of the form

$$\frac{\partial}{\partial t} \begin{pmatrix} f(\mathbf{x}, t) \\ \frac{\partial f(\mathbf{x}, t)}{\partial t} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ c_0\, \nabla^2 & -c_1\, \sqrt{-\nabla^2} \end{pmatrix} \begin{pmatrix} f(\mathbf{x}, t) \\ \frac{\partial f(\mathbf{x}, t)}{\partial t} \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} w(\mathbf{x}, t), \tag{43}$$

where $c_0 = 2^{1/2}\ (a/b)^2$ and $c_1 = 2^{5/4} \cos(\pi/8)\, (a/b)$ are constants.

To compute approximation to the covariance function with scalar x, let's approximate the operators with their Dirichlet counterparts on finite interval $[-L, L]$. Consider the eigenvalue problem

$$-\nabla^2 v_n(x) = -\frac{\partial^2 v_n(x)}{\partial x^2} = \lambda_n^2\, v_n(x), \qquad v_n(-L) = v_n(L) = 0, \tag{44}$$

The normalized (squared) eigenvalues and orthonormal eigenfunctions for $n = 1, 2, \ldots$ are:

$$\begin{aligned} \lambda_n &= \frac{n\,\pi}{2L} \\ v_n(x) &= \sqrt{\frac{1}{L}}\ \sin\left(\frac{n\,\pi\,(x + L)}{2L}\right). \end{aligned} \tag{45}$$

Thus the 1d Laplacian can be associated with the formal kernel

$$K_0(x,x') = -\sum_n \lambda_n^2 \, v_n(x) \, v_n(x'), \tag{46}$$

such that

$$\nabla^2 f(x,t) = \int K_0(x,x') \, f(x,t) \, dx \tag{47}$$

If we expand $f(x,t)$ on the basis $\{v_n(x)\}$ then we have

$$f(x,t) = \sum_n f_n(t) \, v_n(x). \tag{48}$$

where $f_n(t) = \int f(x,t) \, v_n(x) \, dx$. Thus

$$\begin{aligned}
\nabla^2 f(x,t) &= \int K_0(x,x') \, f(x,t) \, dx \\
&= -\sum_{n,n'} \lambda_n^2 \, v_n(x) \, v_n(x') \, f_{n'}(t) \, v_{n'}(x) \, dx \\
&= -\sum_{n,n'} \lambda_n^2 \, v_n(x) \, \delta_{n,n'} \, f_{n'}(t) \\
&= -\sum_n \lambda_n^2 \, v_n(x) \, f_n(t).
\end{aligned} \tag{49}$$

The square root operator $\sqrt{-\nabla^2}$ now has the formal kernel

$$K_1(x,x') = \sum_n \lambda_n \, v_n(x) \, v_n(x'). \tag{50}$$

We can now form (random) series expansion for $w(x,t)$ as follows:

$$\begin{aligned}
w(x,t) &= \sum_n w_n(t) \, v_n(x) \\
w_n(t) &= \int w(x,t) \, v_n(x) \, dx.
\end{aligned} \tag{51}$$

The differential equation can now be expressed in terms of the basis coefficients as follows:

$$\frac{d}{dt} \begin{pmatrix} f_n(t) \\ \frac{df_n(t)}{dt} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -c_0 \, \lambda_n^2 & -c_1 \, \lambda_n \end{pmatrix} \begin{pmatrix} f_n(t) \\ \frac{df_n(t)}{dt} \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} w_n(t). \tag{52}$$

which should be true for all n. The joint spectral density $\tilde{\mathbf{Q}}$ for the process noise can be derived as follows:

$$\begin{aligned}
\mathrm{E}[w_n(t) \, w_m(s)] &= \mathrm{E}[\iint w(x,t) \, v_n(x) \, w(x',s) \, v_m(x') \, dx \, dx'] \\
&= \iint v_n(x) \; \mathrm{E}[w(x,t) \, w(x',s)] \, v_m(x') \, dx \, dx' \\
&= \iint v_n(x) \, Q_c(x-x') \, v_m(x') \, dx \, dx' \, \delta(t-s).
\end{aligned} \tag{53}$$

i.e.,

$$\tilde{Q}_{nm} = \iint v_n(x)\,\mathbf{L}\,Q_c(x-x')\,\mathbf{L}^T\,v_m(x')\,dx\,dx'. \tag{54}$$

with $\mathbf{L} = (0,1)$. Thus we have a model of the form

$$d\mathbf{f} = \mathbf{A}\,\mathbf{f}\,dt + d\mathbf{W}, \tag{55}$$

where $\mathbf{f} = (f_1, df_1/dt, f_2, df_2/dt, \ldots,)$ and

$$\mathbf{A} = \begin{pmatrix} \begin{pmatrix} 0 & 1 \\ -c_0\,\lambda_1^2 & -c_1\,\lambda_1 \end{pmatrix} & & \\ & \begin{pmatrix} 0 & 1 \\ -c_0\,\lambda_2^2 & -c_1\,\lambda_2 \end{pmatrix} & \\ & & \ddots \end{pmatrix} \tag{56}$$

and the diffusion matrix of $\mathbf{W}$ is $\tilde{\mathbf{Q}}$. The measurement model is then

$$y_k = \tilde{\mathbf{H}}_k\,\mathbf{f} + e_k, \tag{57}$$

where $\tilde{\mathbf{H}}_k = (v_1(x_k)\ 0\ v_2(x_k)\ 0\ \cdots)$.

The equation for the mean $\mathbf{m}$ and covariance $\mathbf{P}$ of $\mathbf{f}$ are now given as

$$\frac{d\mathbf{m}}{dt} = \mathbf{A}\,\mathbf{m} \tag{58}$$

$$\frac{d\mathbf{P}}{dt} = \mathbf{A}\,\mathbf{P} + \mathbf{P}\,\mathbf{A}^T + \tilde{\mathbf{Q}}. \tag{59}$$

Let $\mathbf{P}_\infty$ be the solution to the equation

$$\mathbf{A}\,\mathbf{P}_\infty + \mathbf{P}_\infty\,\mathbf{A}^T + \tilde{\mathbf{Q}} = 0 \tag{60}$$

Then we have

$$\mathbf{C}_\mathbf{f}(\tau) = \mathrm{E}[\mathbf{f}(t)\,\mathbf{f}^T(t+\tau)] = \begin{cases} \mathbf{P}_\infty\,\exp(\tau\,\mathbf{A})^T & , \quad \text{for } \tau \geq 0 \\ \exp(-\tau\,\mathbf{A})\,\mathbf{P}_\infty & , \quad \text{for } \tau < 0 \end{cases} \tag{61}$$

where

$$\exp(\tau\,\mathbf{A}) = \begin{pmatrix} \exp\left\{\begin{pmatrix} 0 & 1 \\ -c_0\,\lambda_1^2 & -c_1\,\lambda_1 \end{pmatrix}\tau\right\} & & \\ & \exp\left\{\begin{pmatrix} 0 & 1 \\ -c_0\,\lambda_2^2 & -c_1\,\lambda_2 \end{pmatrix}\tau\right\} & \\ & & \ddots \end{pmatrix} \tag{62}$$

If we define $\mathbf{v}(x) = (v_1(x), v_2(x), \ldots)$, then we have

$$f(x,t) = \sum_n f_n(t)\,v_n(x) = \mathbf{v}^T(x)\,\mathbf{H}\,\mathbf{f}(t) \tag{63}$$

where $\mathbf{H}$ is a matrix with elements $H_{j,2j} = 1$ and thus

$$\begin{aligned} \mathrm{E}[f(x,t)\, f(x+\xi,t+\tau)] &= \mathrm{E}[\mathbf{v}^T(x)\,\mathbf{H}\,\mathbf{f}(t)\,\mathbf{v}^T(x+\xi)\,\mathbf{H}\,\mathbf{f}(t+\tau)] \\ &= \mathbf{v}^T(x)\,\mathbf{H}\;\mathrm{E}[\mathbf{f}(t)\,\mathbf{f}(t+\tau)]\,\mathbf{H}^T\;\mathbf{v}(x+\xi) \\ &= \mathbf{v}^T(x)\,\mathbf{H}\,\mathbf{C}_\mathbf{f}(\tau)\mathbf{H}^T\;\mathbf{v}(x+\xi). \end{aligned} \tag{64}$$

Thus we can approximate the covariance function defined by the stochastic equation by

$$C_f(x,t) \approx \mathbf{v}^T(0)\,\mathbf{H}\,\mathbf{C}_\mathbf{f}(t)\,\mathbf{H}^T\;\mathbf{v}(x). \tag{65}$$

The covariance function can be now numerically computed by using a finite number of terms from this expansion. The Kalman filtering and RTS smoothing based estimation solution can be done by using a finite number of series terms in dynamic model (55) and measurement model (57).

4 Details of Modeling US Monthly Precipitation and Temperature Data

4.1 Model

We implemented the separable spatio-temporal GPs as finite-dimensional SDEs of form as

$$d\mathbf{f}(t) = \mathbf{A}\,\mathbf{f}(t)\,dt + \mathbf{L}\,d\mathbf{W}(t), \tag{66}$$

where matrix $\mathbf{A}$ is a $dN \times dN$ block diagonal matrix, where the $N \times N$ blocks are constructed in such a way that they determine the desired temporal covariance function $C_t(t)$ for the n components (see Hartikainen and Särkkä, 2010, for more details). In this example we used the Matérn temporal covariance model. For the spatial covariance $C_x(\mathbf{x})$ we used 2-dimensional Matérn covariance ($\nu = 3/2$), which is used in forming the elements of diffusion matrix $\mathbf{Q}_c$ of $\mathbf{W}(t)$.

To further lighten up the computations we formed the finite-dimensional model (66) to a latent *inducing process* $\mathbf{u}(t)$ on fixed spatial locations $\{\mathbf{x}_u^i\}_{i=1}^m$, and constructed a linear-Gaussian mapping from the inducing process to a infinite-dimensional latent process as $f(\mathbf{x},t)|\mathbf{u}(t) \sim N(\mathbf{H}(\mathbf{x})\mathbf{u}(t), \mathbf{R}(\mathbf{x})$, where matrices in the mapping are set to $\mathbf{H}(\mathbf{x}) = \mathbf{C}_{\mathbf{f},\mathbf{u}}\mathbf{C}_{\mathbf{u},\mathbf{u}}^{-1}$ and $\mathbf{R}(\mathbf{x}) = \mathrm{diag}(\mathbf{C}_{\mathbf{f},\mathbf{f}} - \mathbf{C}_{\mathbf{x},\mathbf{u}}\,\mathbf{C}_{\mathbf{u},\mathbf{u}}^{-1}\,\mathbf{C}_{\mathbf{u},\mathbf{x}})$, where the covariance terms are evaluated with the spatial covariance function C_x. This can be seen as dynamic formulation of *fully independent conditional* (FIC) sparse approximation recently proposed in the standard GP regression framework. Different approximations can be constructed by choosing the matrices $\mathbf{H}$ and $\mathbf{R}$ appropriately.

To achieve the computational efficiency (i.e., $\mathcal{O}(dm^2)$ complexity in measurement updates) with the low-rank model one can use the matrix inversion lemma to avoid the inversion of $n \times n$ matrix and rather invert a $m \times m$ matrix. In Kalman filtering context the matrix inversion lemma is commonly implemented such that the estimated states and covariances are replaced with *information vectors* and *information matrices*, which are defined as $\mathbf{I}_k = \mathbf{P}_k^{-1}$ and

$\mathbf{i}_k = \mathbf{P}_k^{-1}\mathbf{m}_k$. This formulation is Kalman filter is commonly termed as *information filter* (Grewal and Andrews, 2001). In addition to computational efficiency the information filter is more numerically robust with the low-rank model, which is particularly important in marginal likelihood based hyperparameter learning.

4.2 Data

The data we consider in the paper consists of monthly precipitation and temperature minimum/maximum measurements [1] collected in the US from years 1895-1997. There are 11918 measurements stations for the precipitation data and 8125 for the temperatures. Subsets of this data were used by Paciorek and Schervish (2006) and Vanhatalo and Vehtari (2008) to assess spatial regression models. High fraction of the measurements is missing, and our aim is to fill out the missing measurements by taking account of the spatio-temporal correlations in the data. As the size of original data is very large we focus on (roughly) the same subset of data as in Paciorek and Schervish (2006). The subset is collected from a rectangular area ($[-109.5, -101] \times [36.5, 41.5]$ lon/lat) around Colorado and comprises of 502 stations for the precipitation and 423 for the temperature readings. The total number of measurements in the subset are 372873 for precipitation, 336156 for maximum temperature and 336720 for minimum temperature.

Locations of the measurements stations for precipitation data are shown in Figure 1. Examples of time-series of each data set are shown in Figure 1. The time dynamics of precipitation are much more chaotic than the naturally periodic behavior of temperature readings.

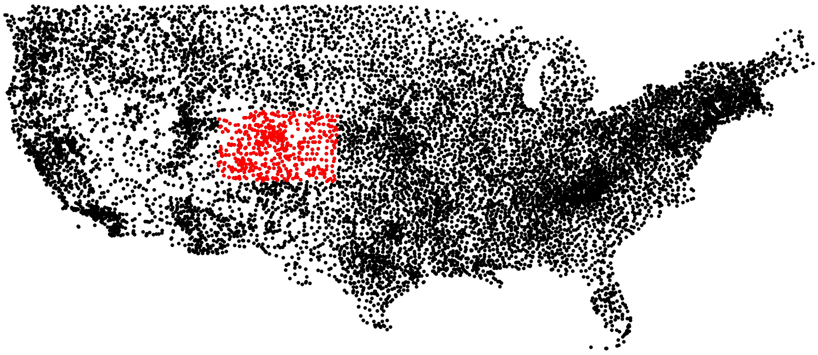

Figure 1: Locations of the measurement stations in the precipitation data. Black dots represent the locations in the whole data, and red dots the locations in the subsample, which used in the experiments. Plots with temperature data are similar, but the number of stations is smaller.

[1] http://www.image.ucar.edu/GSP/Data/US.monthly.met/

References

Cressie, N. and Huang, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *JASA*, 94(448):1330–1340.

Grewal, M. S. and Andrews, A. P. (2001). *Kalman Filtering, Theory and Practice Using MATLAB*. Wiley Interscience.

Hartikainen, J. and Särkkä, S. (2010). Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *Proceedings of MLSP*.

Jazwinski, A. (1970). *Stochastic Processes and Filtering Theory*. Academic Press.

Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*. Springer.

Øksendal, B. (2003). *Stochastic Differential Equations: An Introduction with Applications*. Springer, 6th edition.

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

Råde, L. and Westergren, B. (2004). *Mathematics Handbook*. Studentlitteratur, 5th edition.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Vanhatalo, J. and Vehtari, A. (2008). Modelling local and global phenomena with sparse Gaussian processes. In *Proceedings of UAI*.

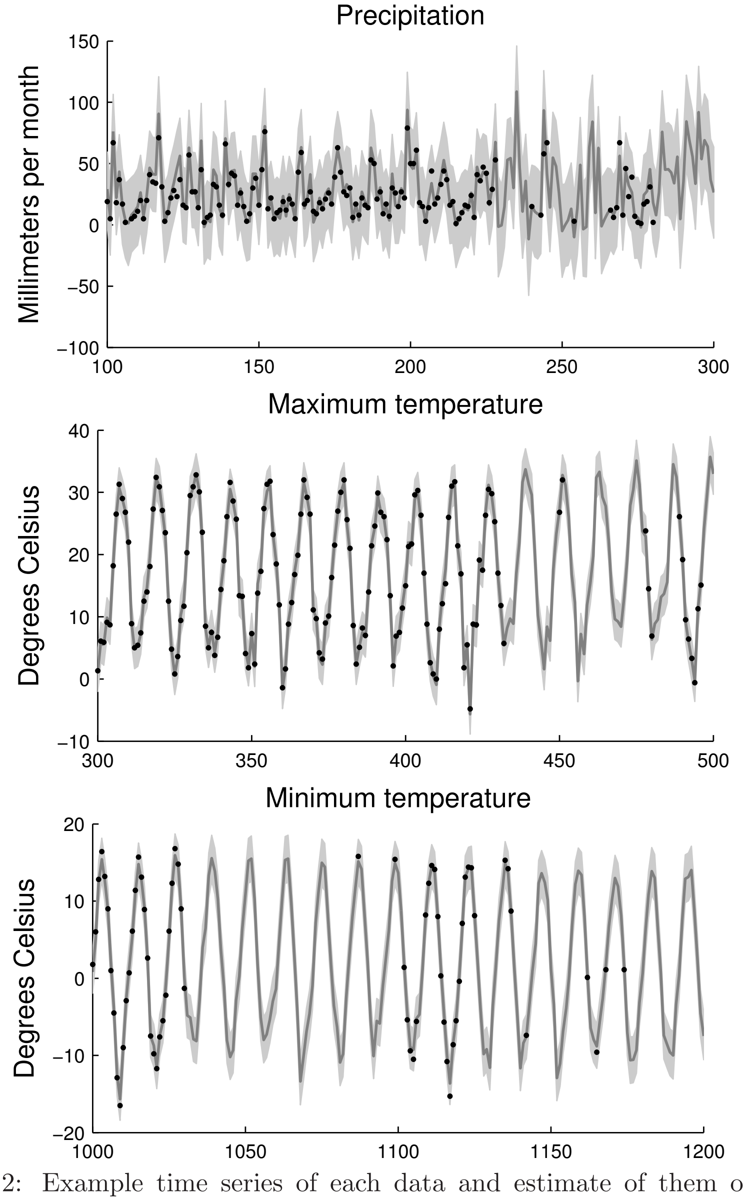

Figure 2: Example time series of each data and estimate of them obtained with STGP ($\nu = 3/2$). Black dots are the measurements, dark gray the mean estimate and light gray the 95% uncertainty.