# Understanding the stochastic partial differential equation approach to smoothing

David L Miller [*] Richard Glennie[*] Andrew E Seaton[*]

**Abstract**

Correlation and smoothness are terms used to describe a wide variety of random quantities. In time, space, and many other domains, they both imply the same idea: quantities that occur closer together are more similar than those further apart. Two popular statistical models that represent this idea are basis-penalty smoothers (Wood, 2017) and stochastic partial differential equations (SPDE) (Lindgren et al., 2011). In this paper, we discuss how the SPDE can be interpreted as a smoothing penalty and can be fitted using the `R` package `mgcv`, allowing practitioners with existing knowledge of smoothing penalties to better understand the implementation and theory behind the SPDE approach.

## 1 Introduction

Data collected over space or time are often obtained with the desire to elicit an underlying pattern. The stochastic partial differential equation (SPDE) approach introduced by Lindgren et al. (2011) and implemented in the `R-INLA` software package (Rue et al., 2009) is a flexible, efficient method to analyse such data. Despite this, wider application is inhibited

[*]Joint first author. *Address:* Centre for Research into Ecological & Environmental Modelling and School of Mathematics & Statistics, University of St Andrews, St Andrews, Fife, Scotland

by two obstacles. First, the methods are presented using mathematical concepts and terms more usually found in applied mathematics and physics, making it difficult for practitioners in other fields to understand and adapt these methods to their own needs. Second, available software implementations are difficult to customise without high-level technical knowledge, limiting application to only those models available in the software or specially requested from software developers.

Here, we aim to mitigate these two issues: we describe (*i*) how the SPDE model can be interpreted as a basis-penalty smoother, a modelling framework more familiar to practitioners who use smoothing techniques (Wood, 2017), and (*ii*) how software to fit these smoothers (e.g., `mgcv`), regularly extended and customised for application, can be used to fit SPDE models or, to go further, used to incorporate SPDE methods into larger models.

In this paper, we consider the following situation. Let $z(x)$ be a random variable observed at location $x$ or time $x$, depending on the domain. A statistical model for $z$ is constructed in three components or terms: $z(x) = \eta(x) + f(x) + \epsilon(x)$. The first component, $\eta(x)$, is the fixed effect, often a linear combination of observed covariates with unknown parameters. The third component, $\epsilon(x)$, represents the measurement error or unstructured error, often $\epsilon(x) \sim \mathcal{N}(0, \sigma^2)$ for unknown parameter $\sigma$ and every location $x$. The second component is a stochastic process, representing the structured dependence among observations: observations made closer together in time or space are more likely to be similar than those further apart. A mathematically convenient and flexible model for this component is a Gaussian process (GP) with mean zero and covariance function $c(x_i, x_j) = \text{Cov}\{f(x_i), f(x_j)\}$. The covariance function quantifies how related two values of $f$ are at two locations. For fixed locations $\{x_1, \ldots, x_n\}$, the value of the GP at these locations, $\{f(x_1), \ldots, f(x_n)\}$, are multivariate Gaussian with mean zero and covariance matrix $\mathbf{\Sigma}$ with $(i, j)^{\text{th}}$ entry $c(x_i, x_j)$. We can extend this formulation to non-Gaussian responses by using a link function, $g$, so the response is then modelled with a specified distribution and a mean $g^{-1}(\eta(x) + f(x))$.

An example of this kind of data might be a time series of counts. The left panel of

Figure 1 show human cases of campylobacterosis (a common form of food poisoning, often originating in under-cooked poultry) in northern Québec, every 28 days from 1 January 1990 to 31 October 2000. We may expect a given time period's count to be similar to its neighbours (e.g. due to seasonal variation), so our aim is to build a model that can capture this dependence. Using the above formulation, we model the counts as Poisson $z(x) \sim \text{Po}(\exp(f(x)))$ where $x$ represents time, $z(x)$ is the number of cases at time $x$, $f$ is a function of time representing the underlying process and exp is the appropriate inverse link function. Dependence structures become more complex when we move to a spatial domain. The right panel of Figure 1 shows remotely-sensed log chlorophyll A levels in the Aral sea, derived from satellite data. In this case we expect that pixels close to each other have similar chlorophyll A levels. We now have $x$ represent a location in space, and the log chlorophyll A level at that location, $z(x)$ is modelled by $z(x) = f(x) + \epsilon(x)$, where now $f$ is a 2-dimensional stochastic process and $\epsilon(x) \sim N(0, \sigma^2)$. We revisit these examples in Section 4, below.

[Figure 1 about here.]

Including $f$ in the model raises two issues: how to specify the covariance function $c$ and, once specified, how to fit the model. There are many possible solutions to these questions, including the SPDE and smoothing penalty approaches, and each uses different theoretical and numerical approximations; however, there is a common element: for observation locations $\{x_1, \ldots, x_n\}$, each method aims to define the covariance between these locations, by constructing an approximation to the precision matrix, defined as the inverse of the covariance matrix ($\mathbf{Q} = \mathbf{\Sigma}^{-1}$). The precision matrix is a fundamental quantity required to fit these models (Simpson et al., 2012). The size of $\mathbf{\Sigma}$ and $\mathbf{Q}$ makes the necessary computations expensive, in particular, if one of these matrices needs to be inverted so as to compute the other. It is in trying to avoid this computational burden that approximations are used.

The SPDE approach provides a method to approximate $\mathbf{Q}$ without the common substantial computational burden. The SPDE is an equation to be solved. Solutions to this

equation are stochastic processes whose covariance structure is chosen to satisfy the relationship the SPDE specifies. The SPDE approach involves finding an SPDE whose solutions have the covariance structure, and implied precision matrix, that is desired for $f$. Lindgren et al. (2011) show how to find an approximate solution to the SPDE by representing $f$ as a sum of basis functions multiplied by coefficients; this provides a computationally efficient way to compute $\mathbf{Q}$: Lindgren et al. (2011) show that the coefficients of these basis functions form a Gaussian Markov random field, for which methods for fast computation of precision matrices already exist (Rue and Held, 2005). These computations make it possible to fit models quickly using integrated nested Laplace approximations (INLA; Rue et al., 2009).

Rue et al. (2017) is a comprehensive review of INLA that defines the class of latent Gaussian models with additive linear predictors which INLA is designed to fit. They also introduce Gaussian Markov random fields and their properties that lead to efficient computation, a key feature of the SPDE approach. Bakka et al. (2018) review the use of INLA for spatial modelling with a focus on the SPDE approach. They give an intuition for the method by introducing the notion of a "discretised" differential operator and describing the finite element methods that are used to solve the SPDE (Brenner and Scott, 2007; Bakka, 2018). See also Krainski et al. (2019) for a collection of worked examples of modelling with SPDEs using `R-INLA`, Wood (2019) for an approach to nested Laplace approximations without sparse Gaussian Markov random field structures, and Blangiardo and Cameletti (2015) for a comprehensive textbook on spatio-temporal modelling with `R-INLA`. We note that the `R-INLA` implementation of the SPDE approach has been applied in a wide variety of domains such as spatial epidemiology (Arab, 2015), species distribution mapping (de Rivera et al., 2018), spatial point processes (Simpson et al., 2016; Yuan et al., 2017; Soriano-Redondo et al., 2019) and environmental science (Huang et al., 2017), to name just a few examples. Our presentation here differs from the above resources in that we explicitly draw links with another well-known modelling framework.

The basis-penalty smoothing approach (Wood, 2017) is similar to the SPDE approach:

the function $f$ is a sum of basis functions multiplied by coefficients. Rather than specify an SPDE and deduce a covariance structure between the coefficients, a smoothing penalty is used to induce correlation between the coefficients. This penalty measures how smooth $f$ is in its domain; intuitively, if $f$ changes more smoothly then values of $f$ at nearby locations are more correlated. Jointly optimising a measure of fit (sum of squares or log-likelihood) and smoothing penalty leads to an optimal curve, the smoothing spline (Wahba, 1990). This is a well-established approach with several excellent introductory resources (Hastie and Tibshirani, 1990; Ramsay and Silverman, 2005; Wood, 2017) and has been applied in many spatio-temporal modelling contexts (recent examples include Wood et al. (2017); Simpson (2018); Pedersen et al. (2019))

There is a direct correspondence between smoothing splines and stochastic processes (Kimeldorf and Wahba, 1970): the smoothing spline is a minimum variance unbiased linear estimator of the posterior mean of the stochastic process. For a stochastic process with a given covariance function, there is a corresponding SPDE and smoothing penalty such that one can estimate the posterior mean of $f$ using the SPDE approach or the basis-penalty approach: both methods estimate the same quantity with the only differences being in numerical approximations and terminology. This means that the SPDE can be interpreted as a smoothing penalty and vice-versa.

This equivalence has been confirmed by Fahrmeir and Lang (2001), Lindgren and Rue (2008), and Yue et al. (2014) who show how basis-penalty smoothers in a Bayesian framework can be interpreted within the SPDE paradigm. Simpson et al. (2012) remark that the SPDE formulation is useful because it provides those with a background in physics or applied mathematics a way to understand and apply the model. In contrast, less emphasis has been placed on discussing this equivalence the other way around: SPDE methods can be formulated as basis-penalty smoothers. The SPDE formulation can seem opaque and fundamentally different for those unfamiliar with the mathematical concepts used. For this reason, showing the approach within the familiar smoothing framework demystifies the workings of

the model and allows researchers in other fields to understand, adapt, and use the methods. We note that our approach is aligned with the general aim of emphasising links between Gaussian processes and the reproducing kernel Hilbert spaces theory that underpins the basis-penalty smoothing approach (Kanagawa et al., 2018), although here we take a more applied perspective.

Our aim in this paper is to show that the SPDE model as introduced by Lindgren et al. (2011) (usually fitted using `R-INLA`) can be described as a basis-penalty smoother and fitted using `mgcv`. To do this, we first describe the SPDE method for those unfamiliar with the mathematical concepts used, highlighting the key steps in the method. Afterward, we show the equivalences and differences between the SPDE method and the analogous basis-penalty smoother.

# 2 The SPDE approach

## 2.1 What is an SPDE?

A stochastic partial differential equation involves stochastic processes and differential operators. Examples of differential operators ($D$) are the first derivative, the second derivative, the gradient operator in two-dimensions or the Laplacian in two dimensions. Combinations of these are also differential operators, e.g., $D = \mathrm{d}/\mathrm{d}x + \mathrm{d}^2/\mathrm{d}x^2$ such that $Df = \mathrm{d}f/\mathrm{d}x + \mathrm{d}^2 f/\mathrm{d}x^2$ for a function $f$. Here, we consider only linear differential operators, that is, $Df$ is a linear combination of derivatives of $f$, of different orders. Differential operators of stochastic processes can be treated similarly to those applied to ordinary functions, there is one key difference that we will highlight below. Overall, an SPDE states that the differential of a function $f$ is equal to some known stochastic process, most commonly the white noise process, $\epsilon$. The white noise process is completely uncorrelated and $\epsilon(x)$ is a Normal random variable with mean zero and finite variance for every $x$.

In general, the SPDE states that $Df = \epsilon$ for some differential operator $D$. A stochastic

process, $f$, is called a solution to the SPDE if it satisfies this equation. Consider an example, let $D$ be the first derivative of the function. The SPDE $Df = \epsilon$ therefore states that the first derivative of $f$ has mean zero and finite variance at every point; furthermore, it states that the value of the derivative at points $x$ and $y$ are uncorrelated for all $x \neq y$. Approximately, this means that for a small $\delta$ and point $x$, $f(x + \delta) = f(x) + \xi$ where $\xi$ is a Normal random variable. Consider if the SPDE has a parameter $\tau$ such that $Df = \epsilon/\tau$ such that $\tau$ controls the variance in the white noise process. This means that changes in $f$ are more variable when $\tau$ is reduced and less variable for higher $\tau$. In other words, the parameters of the SPDE control the smoothness of $f$. It is important to note that here the term "smoothness" is not used in a mathematical way, meaning differentiability, nor in a strictly statistical way, referring to correlation range, but in a qualitative way—when we speak of differentiability or correlation we shall use those terms explicitly.

For a given $D$, the mathematical form of the solution to the SPDE $Df = \epsilon$ is known: $f(x) = \int w(x - u)\epsilon(u)\ \mathrm{d}u$ where $w$ is a function you can derive given you know $D$. The function $w$ is called Green's function; in the appendix (Proposition 1) we show how this function is derived from $D$. Intuitively, $w$ acts as a weighting function such that the value of the stochastic process at $x$ is a weighted sum over the white noise process; this is called a convolution. Suppose $w$ were set to give infinite weight to distance 0, $w(0) = \infty$, and zero elsewhere, $w(d) = 0$ for $d \neq 0$, then $f(x) = \epsilon(x)$: $f$ is just the white noise process, completely uncorrelated. Alternatively, if $w$ gave equal weight to all distances, e.g., $w(d) = 1$ for all $d$, then $f(x)$ would be constant, perfectly correlated. Between these two extremes are weighting functions that reproduce correlations over different ranges. It can be shown that the covariance function is given by $c(x, y) = \int w(x - u)w(y - u)\ \mathrm{d}u$, see appendix (Proposition 2) for the derivation.

In summary, the solutions to the SPDE $Df = \epsilon$ have a covariance structure that is induced by the choice of $D$. This means that one could describe a system using an SPDE and then deduce the associated covariance function from it. The power of the SPDE approach

is realised by doing the opposite: find a $D$ that induces the covariance function that you want. The power of finding the SPDE corresponding to a desired covariance function is that the precision matrix can be efficiently computed using the SPDE.

## 2.2 Solving the SPDE

The SPDE involves applying a differential operator $D$ to a stochastic process, $f$, but this cannot be done in the same way as when you apply $D$ to a known function. This is because $f$ is random and, in many cases, realisations of $f$ will not be suitably differentiable. For example, the Brownian motion stochastic process has a derivative equal to the white noise process, but it is also known that simulated trajectories of Brownian motion are nowhere differentiable. $Df = \epsilon$ is a convenient shorthand way to think about the SPDE, but technically, the SPDE only has meaning when stated in an integral form. That is, $Df = \epsilon$ means that we require $\int Df(x)\phi(x)\ \mathrm{d}x = \int \epsilon(x)\phi(x)\ \mathrm{d}x$ for every function $\phi$ with compact support. The function $\phi$ is often called the test function. For brevity, let $\langle f, g\rangle = \int f(x)g(x)\ \mathrm{d}x$ and so the integral form is $\langle Df, \phi\rangle = \langle \epsilon, \phi\rangle$. The notation $\langle f, g\rangle$ is called the inner product of $f, g$, it has many nice mathematical properties, including being linear, that is $\langle \sum_{i=1}^{n} a_i f_i, \sum_{j=1}^{m} b_i g_i\rangle = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \langle f_i, g_j\rangle$ for functions $f_1, \ldots, f_n, g_1, \ldots, g_m$ and constants $a_1, \ldots, a_n, b_1, \ldots, b_m$.

In the integral form, the equation makes sense because any stochastic process can be integrated, but not every one can be differentiated. By requiring the equation to hold for every $\phi$, we require the left-hand stochastic process $Df$ and the right-hand process $\epsilon$ to have the same integral, no matter how we average over space. For example, if the stochastic processes were one-dimensional, we could split the real line into intervals $[n, n+1]$ and select a function $\phi_n$ to be one on this interval and zero outside. Since the integral equation must hold for all such functions, we therefore require $Df$ to have the same average value as $\epsilon$ on each and every interval.

Given an SPDE, Lindgren et al. (2011) show how to derive an approximate solution using

the finite element method. The domain (e.g., time or space) is split into "elements", e.g., a grid or a triangulation, often called a mesh. To each point $j = 1, \ldots, M$ on this mesh, a basis function $\psi_j$ is associated. The solution to the SPDE is then a weighted sum of the basis functions and random variables $\beta_j$: $f(x) = \sum_{j=1}^{M} \beta_j \psi_j(x)$.

The integral form of the SPDE then implies that for any function $\phi$, $\sum_{j=1}^{M} \beta_j \langle D\psi_j, \phi \rangle = \langle \epsilon, \phi \rangle$. We cannot, however, check this equation for infinitely many test functions $\phi$, so instead we restrict to only testing with the functions that can be written in our chosen basis. As $D$ is a linear operator, this is equivalent to solving the system of equations $\sum_{j=1}^{M} \beta_j \langle D\psi_j, \psi_i \rangle = \langle \epsilon, \psi_i \rangle$ for every $i = 1, \ldots, M$. This system can be written as a matrix equation: $\mathbf{P}\boldsymbol{\beta} = \mathbf{e}$ where $\mathbf{P}$ has $(i,j)^{\text{th}}$ entry $\langle D\psi_i, \psi_j \rangle$ and $\mathbf{e}$ has $j^{\text{th}}$ entry $\langle \epsilon, \psi_j \rangle$.

To summarise, the SPDE is written in an integral form, sometimes using inner products, since stochastic processes are well defined when integrated but not when differentiated. Given this, the solution is represented in a chosen basis. The integral form is then solved by considering only test functions within that basis. This leads to a matrix equation involving the coefficients $\boldsymbol{\beta}$, the matrix $\mathbf{P}$, and the random vector $\mathbf{e}$. The random vector $\mathbf{e}$ has known distribution, because it depends only on the basis functions and the white noise process: $\mathbf{e}$ has a multivariate Gaussian distribution with mean zero and a precision matrix $\mathbf{Q}_e$ where $\mathbf{Q}_e^{-1}$ has $(i,j)^{\text{th}}$ entry $\langle \psi_i, \psi_j \rangle$. It follows from $\mathbf{P}\boldsymbol{\beta} = \mathbf{e}$ that $\boldsymbol{\beta} \sim N(0, \mathbf{Q}^{-1})$ where $\mathbf{Q} = \mathbf{P}^T \mathbf{Q}_e \mathbf{P}$. The SPDE is therefore a way to specify a prior for $\boldsymbol{\beta}$.

This provides an approximate solution to the SPDE. For example, given an SPDE, one can use the finite element method to compute $\mathbf{Q}$ and therefore simulate $\tilde{\boldsymbol{\beta}}$ from a multivariate Gaussian distribution with precision $\mathbf{Q}$. The function $\tilde{f} = \sum_{j=1}^{M} \tilde{\beta}_j \psi_j$ would then be a realisation from a stochastic process which is a solution to the SPDE, a stochastic process with the covariance structure implied by $D$.

## 2.3 Matérn SPDE

The focus of Lindgren et al. (2011) and the covariance function most commonly used in the `R-INLA` software is the Matérn covariance function. The Matérn covariance function is considered a flexible model for the dependencies found in real world observations: it has the form $c(x,y) = \frac{2^{1-\nu}}{(4\pi)^{d/2}\kappa^{2\nu}\tau^2\Gamma(\nu+d/2)}(\kappa\|x-y\|)^\nu K_\nu(\kappa\|x-y\|)$ where $\nu, \kappa, \tau$ are parameters, $K_\nu$ is the modified Bessel function of the second kind, and $d$ is the dimension of the domain. Figure 2 shows realisations from two stochastic processes with Matérn covariance functions in one-dimension, one with a longer correlation range than the other.

It is difficult to fit models with this covariance structure due to the computational issues mentioned above. Lindgren et al. (2011) apply the finite element method to approximate stochastic processes with Matérn covariance (a comparison of the notation used in Section 2.2 and that used in Lindgren et al. (2011) is given in the appendix, Section 5). To do this, they present the differential operator that corresponds to this covariance function: $D = (\kappa^2 - \Delta)^{\alpha/2}\tau$ where $\alpha = \nu + d/2$.

When $\alpha \neq 2$, this is called a fractional differential operator; for this paper, we consider only the case when $\alpha = 2$ and so $D$ is again a linear differential operator. In practice, $\alpha$ is poorly identified and difficult to estimate from data, so its value is often assumed to be fixed (Zhang, 2004).

Lindgren et al. (2011) solve the SPDE $\kappa^2 f - \Delta f = \epsilon/\tau$ using the finite element method. By deriving the weighting function and computing the covariance from this SPDE, Whittle (1954) shows that the solutions have Matérn covariance, as desired. In other words, the precision matrix computed from the finite element method is an approximation to the precision matrix one would obtain if you computed the variance-covariance matrix $\mathbf{\Sigma}$ with the Matérn covariance function and then, at great computational cost, inverted this matrix. Figure 2 shows a subview of the approximate precision matrix: the matrix is mostly filled with zeroes, with non-zero values occurring on three bands down the diagonal. This is an example of a sparse matrix, computations with these matrices are efficient because many

of the computations ordinarily required can be omitted as it is known the matrix is mostly zeroes.

To use the finite element method one must chose a mesh, a grid or triangulation, over the domain and a basis to define on this mesh. The default choice in `R-INLA` is to use a regular grid in 1D (or a constrained Delaunay triangulation in 2D) to produce a mesh and then define piecewise linear basis functions (specifically, linear B-spline basis functions) on this mesh.

[Figure 2 about here.]

# 3 The basis-penalty approach

## 3.1 What is a basis-penalty smoother?

The basis-penalty approach refers to models where $f$ is assumed to have the form $f(x) = \sum_{j=1}^{M} \beta_j \psi_j$ for $M$ basis functions $\psi_1, \ldots, \psi_M$ and parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$. A model is then assumed for the observations given this form for $f$ to provide a measure of fit, the log-likelihood $l(\boldsymbol{\beta})$. Alternatively, the sum-of-squares can be used as a measure of fit. Optimising to obtain $\hat{\boldsymbol{\beta}}$ leads to a function $\hat{f}$ that, given $M$ is large enough, will interpolate the observed data: capturing the noise in the observations as well as the underlying signal (such overfitting stops us from making inference on the signal). A smoothing penalty, $J(\boldsymbol{\beta}, \lambda)$, is subtracted from the log-likelihood to penalize functions that are too wiggly. The smoothing parameter, $\lambda$, controls the extent of the penalization (a larger value of $\lambda$ leads to a smoother $\hat{f}$).

The estimates for $\boldsymbol{\beta}$ are defined to be those that optimise the joint measure of fit and smoothness, the penalised likelihood: $l_p(\boldsymbol{\beta}, \lambda) = l(\boldsymbol{\beta}) - J(\boldsymbol{\beta}, \lambda)$. This involves estimating both the optimal smoothing parameter $\lambda$ and coefficients $\boldsymbol{\beta}$. In practice, REstricted Maximum Likelihood (REML; Wood, 2011) is used to to do this.

There are several choices for the smoothing penalty. Most are defined using a differential operator $D$. For example, in one dimension, the smoothing penalty $J(\boldsymbol{\beta}, \lambda) =$

$\lambda \int (\partial^2 f/\partial x^2)^2 \, \mathrm{d}x$, i.e., where $D$ is the second derivative, is often used. For this penalty, functions with rapidly changing gradients are penalised while functions with constant gradient, straight lines, have no penalty. In higher dimensions, the thin-plate spline (Wood, 2003) is often used with penalty: $J(\boldsymbol{\beta}, \lambda) = \lambda \int (\partial^2 f/\partial x^2)^2 + 2\,(\partial^2 f/\partial x \partial y)^2 + (\partial^2 f/\partial y^2)^2 \, \mathrm{d}x\mathrm{d}y$ for two-dimensions. This penalty takes the total variation in the gradient of $f$ including the interaction between the coordinates. The penalty for smoothing splines takes the form $J(\boldsymbol{\beta}, \lambda) = \lambda \int (Df)^2 \, \mathrm{d}x$ for some chosen differential operator $D$ (see Yue and Speckman (2010) and Yue et al. (2014) who show this for the thin plate spline penalty). This can also be written as an inner product $J(\boldsymbol{\beta}, \lambda) = \lambda \langle Df, Df \rangle$.

When $f(x) = \sum_{j=1}^{M} \beta_j \psi_j(x)$, the penalty based on the differential operator $D$ can be written in matrix form: $J(\boldsymbol{\beta}, \lambda) = \lambda \boldsymbol{\beta}^\top \mathbf{S} \boldsymbol{\beta}$ where $\mathbf{S}$ is a $M \times M$ matrix with $(i,j)^{\text{th}}$ entry $\langle D\psi_i, D\psi_j \rangle$.

In summary, a basis-penalty smoother is specified by selecting a basis, e.g., a B-spline basis of specified order, and a smoothing penalty. The parameters are then estimated by optimising the penalised likelihood: $l_p(\boldsymbol{\beta}, \lambda) = l(\boldsymbol{\beta}) - \lambda \boldsymbol{\beta}^\top \mathbf{S} \boldsymbol{\beta}$.

## 3.2 Connection between SPDE and penalty

Rewriting the penalized log-likelihood as a likelihood we obtain $\exp\{l_p(\boldsymbol{\beta}, \lambda)\} = \exp\{l(\boldsymbol{\beta})\} \times \exp(-\lambda \boldsymbol{\beta}^\top \mathbf{S} \boldsymbol{\beta})$. A Bayesian interpretation of the penalised likelihood as proportional to a posterior implies that $\exp(-\lambda \boldsymbol{\beta}^\top \mathbf{S} \boldsymbol{\beta})$ is an improper prior for $\boldsymbol{\beta}$ (Silverman, 1985; Wood, 2017). Since $\exp(-\lambda \boldsymbol{\beta}^\top \mathbf{S} \boldsymbol{\beta})$ is proportional to a multivariate normal distribution with mean zero and precision matrix $\mathbf{S}_\lambda = \lambda \mathbf{S}$, the penalized likelihood is equivalent to assigning the prior $\boldsymbol{\beta} \sim N(0, \mathbf{S}_\lambda^{-1})$.

The connection between the SPDE approach and the basis-penalty approach can now be made clear. It can be shown that for a given differential operator $D$, the approximate precision matrix for the SPDE $Df = \epsilon$ is the same as the precision matrix $S_\lambda$ computed using the smoothing penalty $\langle Df, Df \rangle$ (appendix, Proposition 3).

This connection has two implications. First, it means that the differences between the basis-penalty approach and the SPDE finite element approximation, when using the same basis and differential operator, are differences in implementation only, as both should lead to the same approximate precision matrix. Second, the connection means that any SPDE of the form $Df = \epsilon$ can be understood and interpreted as a smoothing penalty of the form $\langle Df, Df \rangle = \int \{Df(x)\}^2 \, \mathrm{d}x$, and vice-versa.

## 3.3 Matérn penalty

The SPDE specified in Lindgren et al. (2011) has the differential operator $D = \tau(\kappa^2 - \Delta)$. Given the connection described above, this can be interpreted as a smoothing penalty: $\tau^2 \int (\kappa^2 f - \Delta f)^2 \, \mathrm{d}x$. This penalty is different from those considered above because it contains two smoothing parameters: $\tau$ and $\kappa$. This offers it more flexibility. The penalty can still, however, be interpreted as such: it is a trade-off between the value of the function $f$ and the second derivative $\Delta f$ in each direction. As $\kappa$ is increased, the value of $\kappa^2 f$ increases, meaning that $\Delta f$ can be higher, the function be less smooth, while keeping the penalty the same. Alternatively, $\kappa$ can be described as the inverse correlation range: higher values of $\kappa$ lead to less smooth functions meaning values of the function become less correlated. The smoothing parameter $\tau$ controls the overall smoothness of $f$.

The Matérn penalty can be written in matrix form as above, but for computational convenience, it is first broken into three parts: $\langle Df, Df \rangle = \tau^2(\kappa^4 \langle f, f \rangle + 2\kappa^2 \langle \nabla f, \nabla f \rangle + \langle \Delta f, \Delta f \rangle)$. Notice that it appears that the Laplacian $\Delta$ has been replaced with the gradient operator $\nabla$: this relationship holds here using Green's first identity and the Neumann boundary condition, see Bakka et al. (2018) for more detail. This leads to the smoothing matrix $\mathbf{S} = \tau^2(\kappa^4 \boldsymbol{C} + 2\kappa^2 \boldsymbol{G}_1 + \boldsymbol{G}_2)$ where $\boldsymbol{C}, \boldsymbol{G}_1, \boldsymbol{G}_2$ are all $M \times M$ matrices with $(i,j)^{\text{th}}$ entries $\langle \psi_i, \psi_j \rangle$, $\langle \nabla \psi_i, \nabla \psi_j \rangle$, and $\langle \Delta \psi_i, \Delta \psi_j \rangle$, respectively. All of these matrices are sparse and so computation of the smoothing penalty, $\boldsymbol{\beta}^\top \mathbf{S} \boldsymbol{\beta}$, is computationally efficient. The matrix $\mathbf{S}$ is equal to the matrix $\mathbf{Q} = \mathbf{P}^\top \mathbf{Q}_e \mathbf{P}$ computed using the finite element method (Appendix,

Proposition 3).

### 3.4 Fitting the Matérn SPDE in `mgcv`

The `mgcv` R package allows the specification of new basis-penalty smoothers by writing new "`smooth.construct`" functions which build an appropriate design matrix (containing evaluations of the basis functions), penalty matrices and other optional components. Within this framework we can fit the SPDE model in `mgcv` providing a `smooth.construct.spde.smooth.spec` constructor. `R-INLA` provides helper functions to construct the required design and penalty matrices. Here we sketch an algorithm for setting-up SPDE models in `mgcv`.

Given we have a response $\{y_i; i = 1, \ldots, n\}$ and covariates in an $n \times n_c$ matrix $\mathbf{X}_c$, we construct our model as follows.

1. Create a mesh using `INLA::inla.mesh.1d` or `INLA::inla.mesh.2d`.

2. Calculate $\mathbf{C}$, $\mathbf{G}_1$ and $\mathbf{G}_2$ using `INLA::inla.mesh.fem` (`c1`, `g1` and `g2`, respectively).

3. We need to connect the basis representation of $f$ to the observation locations. At present $\boldsymbol{\beta}$ contains the value of $f$ at each mesh point, not at each observation location. A matrix multiplication is used to project the values at all mesh points to the observations locations, it is called the projection matrix $\mathbf{A}$ (found using `INLA::inla.spde.make.A`). The full design matrix $\boldsymbol{X}$ is then given by combining the fixed effects design matrix $\mathbf{X}_c$ and the contribution for $f$, $\mathbf{A}$.

4. Having constructed the design matrix and penalty matrices, use REML to find optimal $\kappa$, $\tau$ and $\boldsymbol{\beta}$ subject to the penalty matrix $\kappa^4\mathbf{C} + 2\kappa^2\mathbf{G}_1 + \mathbf{G}_2$. (Parametrisation for this model in `mgcv` is given in Supplementary Material section 4.)

As REML is an empirical Bayes procedure, we expect point estimates for $\hat{\boldsymbol{\beta}}$ to coincide for the procedure outlined above and `R-INLA`. A uniform prior is implied for the smoothing parameters ($\tau$ or $\kappa$); `R-INLA` allows for similar estimation by just using the modes of the

hyperparameters $\kappa$ and $\tau$ (the `int.strategy="eb"` option). Proper priors could be used if step (4), above, was replaced by an MCMC scheme.

# 4 Examples

We now compare the SPDE and basis-penalty models applied to three example datasets. We fitted the SPDE Matérn model in both `R-INLA` and `mgcv`. Code for these examples is available as supplementary material.

## 4.1 Campylobacterosis cases in Québec

Ferland et al. (2006) analyse a time series of (human) cases of campylobacterosis in northern Québec, with observations every 28 days from 1 January 1990 to 31 October 2000 (140 observations). We modelled the number of infections as a function of time, using a Poisson response and a log link function. The model is fitted using three approaches (*i*) a Matérn basis-penalty smoother with 50 degree 2 B-splines, fitted with `mgcv`; (*ii*) a Matérn SPDE for $f$ with a finite element basis of 50 degree 2 B-splines and penalized complexity priors (Simpson et al., 2017) on smoothing parameters, fitted with `R-INLA`; (*iii*) a basis-penalty smoother with penalty equal to the integral of the squared second derivative, using 50 degree 2 B-splines fitted using `mgcv`.

[Figure 3 about here.]

Fitted models are shown in Figure 3. Results from the `R-INLA` and `mgcv` SPDE implementations are very similar. This is supported by the similarity in the estimated hyperparameters ($\tau = 3.603$ and $\kappa = 0.429$ for `R-INLA`, and $\tau = 3.252$ and $\kappa = 0.475$ for `mgcv`). The squared second derivative penalty B-spline fit from `mgcv` is smoother than those from the SPDE-based methods.

## 4.2 Aral sea chlorophyll

Moving to a 2-dimensional smoothing problem, we consider remotely sensed (log) chlorophyll from the Aral sea collected by the NASA SeaWifs satellite over a series of 8 day observation periods. The 496 observations used here are averages (from 1998 to 2002) of the 38th observation period. Data were taken from the `gamair` package (dataset `aral`) and consist of spatial coordinates and logarithm of chlorophyll concentration.

We built a mesh using `fmesher::meshbuilder` (Supplementary Figure 2) and generated two-dimensional degree 1 B-splines. We consider the model $y_i = f(\boldsymbol{x}_i) + \epsilon$ for location $\boldsymbol{x}_i$ with no fixed effects. We fitted the Matérn model using the SPDE and penalty approaches in `R-INLA` and `mgcv`. For the `R-INLA` model, penalized complexity priors were used.

There was little visual difference with good agreement in the predictions (Supplementary Figure 3 shows the posterior mode and percentile credible surfaces for these models and Supplementary Figure 4 gives a graphical comparison). Hyperparameter estimates were similar: $\tau = 0.059$ and $\kappa = 3.43$ for `R-INLA` and $\tau = 0.059$ and $\kappa = 3.543$ for `mgcv`. To investigate differences between the two models we took (1000) samples from the posterior of each model and looked at the differences between pairs of realisations on a per-cell basis. Plots of the mean of these differences and their standard deviations are shown in Figure 4. The mean plot shows structure to the differences in the models, though differences are relatively small (range of log chlorophyll A values in original data: 1.905–19.275). This is to be expected if the models produce similar predicted values to each other, which are consistent through each realisation.

[Figure 4 about here.]

## 4.3 MODIS land surface temperatures

To compare the two approaches on a large data set, we now turn to land surface temperature data collected by the Terra instrument on the MODIS satellite. The data consist of a

$500 \times 300$ grid of measurements in the latitude range 34.29519–37.06811 and longitude range -95.91153–-91.28381 on August 4, 2016. The training data (105,569 observations) as defined in Heaton et al. (2018) were used to fit the model, but a significantly simpler mesh was used (see Supplementary Figure 5). Following from Heaton et al. (2018) we assumed a Gaussian response for temperature and fitted a 2-dimensional SPDE model on latitude and longitude.

As the dataset was quite large, we used the `bam` ("big additive model") function in `mgcv` to fit the SPDE model (additionally discretizing covariate values for efficient storage and computation; Wood et al., 2017). The SPDE fitted by `R-INLA` used the empirical Bayes integration strategy. We timed the fitting for both approaches (ignoring mesh setup, which was shared across methods), taking only the time for `inla` and `bam` to run. The `mgcv` model was slightly faster (4.71 minutes versus 5.23 in `R-INLA` running on a Windows server using 1 core of a Xeon Gold 6152 at 2.1GHz with 512Gb RAM). Supplementary Figure 6 compares the predictions from the two methods, the largest absolute difference between predictions is 4.761, which is small compared to the range of the data.

# 5 Discussion

We have drawn links between two approaches to fitting the same model: the stochastic partial differential equation method as implemented in `R-INLA` and the basis-penalty smoothing approach as fitted in `mgcv`. This paper aims to make accessible what is equivalent between the approaches, what is a matter of choice, and what is fundamentally different. Yue et al. (2014) show how splines can be specified using the SPDE approach, benefitting those familiar with SPDEs. Here, we do the opposite for the benefit of those familiar with the (penalized likelihood/empirical Bayes) GAM framework. Supplementary Figure 1 is a flow diagram showing the parallels between the smoothness and correlation approaches we have discussed.

Similarities between many smoothing techniques can be drawn. Smoothing splines, kriging, Gaussian Markov random fields, and SPDEs approximate similar models, but their

explanations make it difficult for practitioners to appreciate their commonalities and determine precisely what is a necessary and what is a coincidental association. Taking the precision matrix as the common currency between these methods, a modelling framework emerges:

1. **Choose a covariance model**: explicitly, as in kriging, through the smoothness penalty as with smoothing splines, or with an SPDE;

2. **Approximate the precision matrix $\mathbf{Q}$**: reduce dimension (fixed rank kriging, thin plate splines) or induce sparsity in $\mathbf{Q}$ (B-splines, SPDE);

3. **Draw approximate inference using a software implementation**: e.g., with `mgcv`, MCMC (e.g., Stan (Carpenter et al., 2017); `JAGS`, (Plummer, 2017)), `R-INLA` (Rue et al., 2009), `lme4` (Bates et al., 2015) or `TMB` (Kristensen et al., 2016).

This paper is an example of comparing two methods according to this framework. Doing so for other smoothing methods will allow alternative modelling approaches to be compared on the grounds of genuine differences: in the covariance function, in the approximation for $\mathbf{Q}$, in the estimation procedure, or, simply, in the software implementation.

# Acknowledgements

The authors wish to thank the editor and two anonymous reviewers for their constructive comments on the first submission of the manuscript. They also thank Finn Lindgren for extremely helpful input and Simon Wood for the suggestion of the smoothing parameter parameterisation. The authors also thank Steve Buckland, David Borchers and Joe Watson for their comments on the manuscript.

# References


Arab, A. (2015). Spatial and spatio-temporal models for modeling epidemiological data with excess zeros. *International Journal of Environmental Research and Public Health 12*(9), 10536–10548.

Bakka, H. (2018). How to solve the stochastic partial differential equation that gives a Matérn random field using the finite element method. *arXiv preprint arXiv:1803.03765*.

Bakka, H., H. Rue, G.-A. Fuglstad, A. Riebler, D. Bolin, J. Illian, E. Krainski, D. Simpson, and F. Lindgren (2018). Spatial modeling with r-inla: A review. *Wiley Interdisciplinary Reviews: Computational Statistics 10*(6), e1443.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software 67*(1).

Blangiardo, M. and M. Cameletti (2015). *Spatial and Spatio-temporal Bayesian Models with R-INLA*. Wiley, Chichester, UK.

Brenner, S. and R. Scott (2007). *The mathematical theory of finite element methods*, Volume 15. Springer, New York, NY, USA.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software 76*(1), 1–32.

de Rivera, O. R., M. Blangiardo, A. López-Quílez, and I. Martín-Sanz (2018). Species distribution modelling through Bayesian hierarchical approach. *Theoretical Ecology*, 1–11.

Fahrmeir, L. and S. Lang (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 50*(2), 201–220.

Ferland, R., A. Latour, and D. Oraichi (2006, November). Integer-Valued GARCH Process. *Journal of Time Series Analysis 27*(6), 923–942.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Number 43 in Monographs on Statistics and Applied Probability. Chapman and Hall, London, UK.

Heaton, M. J., A. Datta, A. O. Finley, R. Furrer, J. Guinness, R. Guhaniyogi, F. Gerber, R. B. Gramacy, D. Hammerling, M. Katzfuss, F. Lindgren, D. W. Nychka, F. Sun, and A. Zammit-Mangion (2018, December). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*.

Huang, J., B. P. Malone, B. Minasny, A. B. McBratney, and J. Triantafilis (2017). Evaluating a Bayesian modelling approach (INLA-SPDE) for environmental mapping. *Science of the Total Environment 609*, 621–632.

Kanagawa, M., P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur (2018, July). Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences. *arXiv preprint arXiv:1807.02582*. arXiv: 1807.02582.

Kimeldorf, G. S. and G. Wahba (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics 41*(2), 495–502.

Krainski, E., V. Gómez-Rubio, H. Bakka, A. Lenzi, D. Castro-Camilo, D. Simpson, F. Lindgren, and H. Rue (2019). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. CRC Press, Boca Raton, FL, USA.

Kristensen, K., A. Nielsen, C. W. Berg, H. Skaug, and B. M. Bell (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software 70*(5), 1–21.

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

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

Pedersen, E. J., D. L. Miller, G. L. Simpson, and N. Ross (2019). Hierarchical generalized additive models in ecology: an introduction with mgcv. *PeerJ 7*, e6876.

Plummer, M. (2017). *JAGS Version 4.3.0 user manual*.

Ramsay, J. and B. W. Silverman (2005). *Functional Data Analysis* (2 ed.). Springer Series in Statistics. New York: Springer-Verlag.

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

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

Rue, H., A. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson, and F. K. Lindgren (2017). Bayesian computing with inla: A review. *Annual Review of Statistics and Its Application 4*(1), 395–421.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B (Methodological) 47*(1), 1–52.

Simpson, D., J. B. Illian, F. Lindgren, S. H. Sørbye, and H. Rue (2016). Going off grid:

computationally efficient inference for log-Gaussian Cox processes. *Biometrika 103*(1), 49–70.

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

Simpson, D., H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye (2017). Penalising model component complexity: a principled, practical approach to constructing priors. *Statistical Science 32*(1), 1–28.

Simpson, G. L. (2018). Modelling Palaeoecological Time Series Using Generalised Additive Models. *Frontiers in Ecology and Evolution 6*.

Soriano-Redondo, A., C. M. Jones-Todd, S. Bearhop, G. M. Hilton, L. Lock, A. Stanbury, S. C. Votier, and J. B. Illian (2019). Understanding species distribution in dynamic populations: a new approach using spatio-temporal point process models. *Ecography 42*(6), 1092–1102.

Wahba, G. (1990). *Spline models for observational data*. SIAM, Philadephia, PA, USA.

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

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(1), 95–114.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(1), 3–36.

Wood, S. N. (2017). *Generalized Additive Models. An Introduction with R* (2nd ed.). Texts in Statistical Science. CRC Press, Boca Raton, FL, USA.

Wood, S. N. (2019). Simplified integrated nested Laplace approximation. *Biometrika*.

Wood, S. N., Z. Li, G. Shaddick, and N. H. Augustin (2017, July). Generalized Additive Models for Gigadata: Modeling the U.K. Black Smoke Network Daily Data. *Journal of the American Statistical Association 112*(519), 1199–1210.

Yuan, Y., F. E. Bachl, F. Lindgren, D. L. Borchers, J. B. Illian, S. T. Buckland, H. Rue, T. Gerrodette, et al. (2017). Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. *The Annals of Applied Statistics 11*(4), 2270–2297.

Yue, Y. and P. L. Speckman (2010). Nonstationary spatial gaussian markov random fields. *Journal of Computational and Graphical Statistics 19*(1), 96–116.

Yue, Y. R., D. Simpson, F. Lindgren, and H. Rue (2014). Bayesian adaptive smoothing splines using stochastic differential equations. *Bayesian Analysis 9*(2), 397–424.

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

# List of Figures

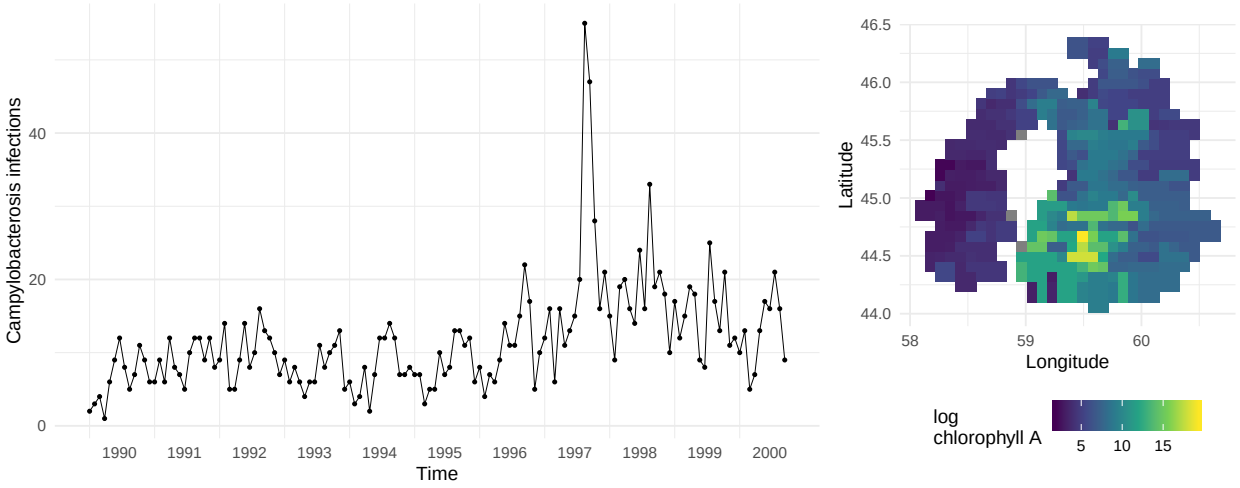


Figure 1: Examples of data with underlying dependence between observations. Left shows counts of campylobacterosis infections in northern Québec, summarized every 28 days from 1 January 1990 to 31 October 2000. Right shows the raw log chlorophyll A in the Aral sea from the SeaWIFS satellite. In both cases we can build a model that takes into account the structure in the data.

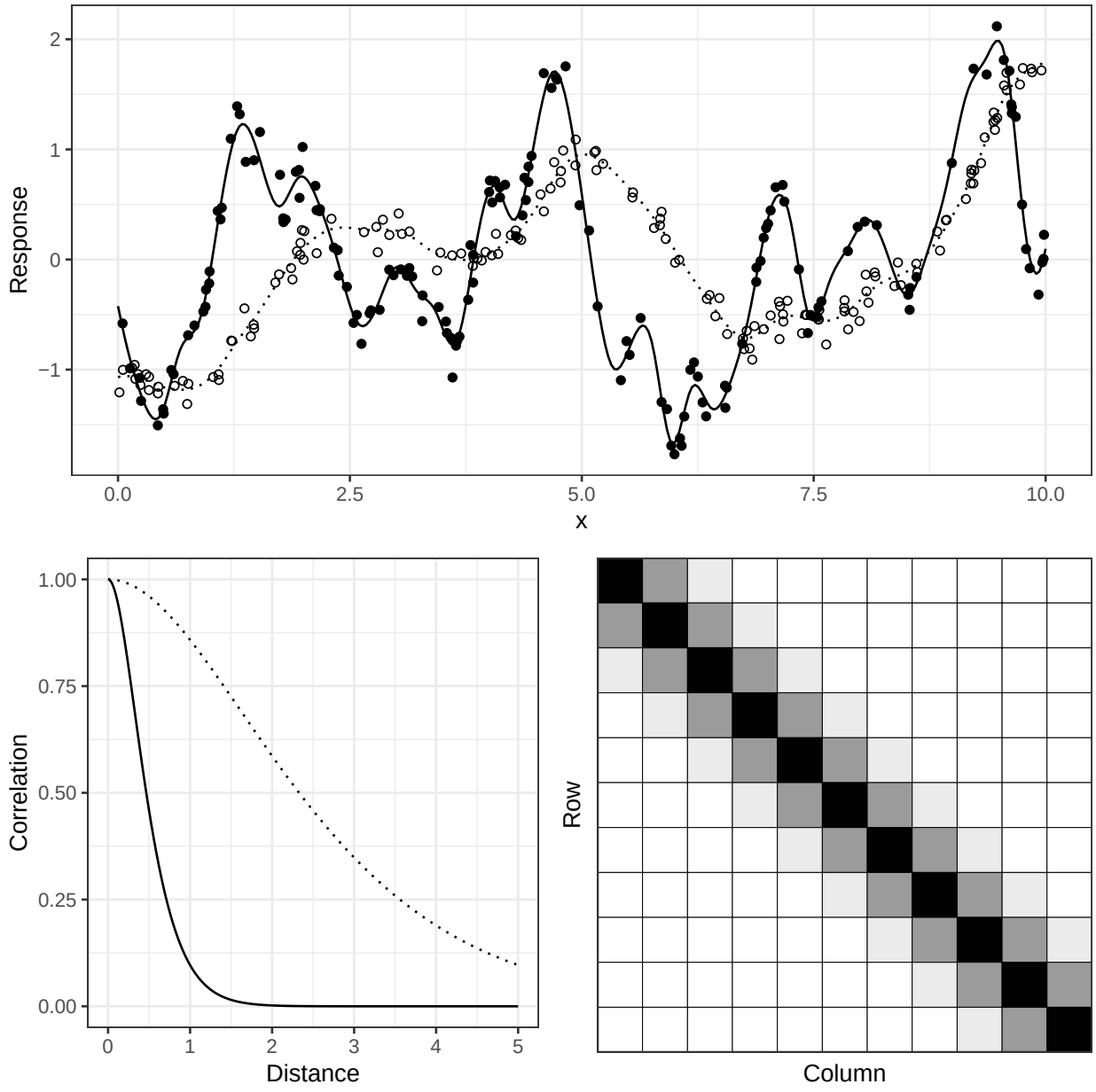


Figure 2: Two functions, one smooth (long-range correlation, dashed line, open circle data) and one rough (short-range correlation, solid line, filled circle data) (top plot), their Matérn correlation functions (bottom left plot, same line types) and the first 11 rows and columns of an example approximate Gaussian Markov Random field precision matrix (bottom right plot, darker shade indicates higher absolute value, each row and column corresponds to a data point location).

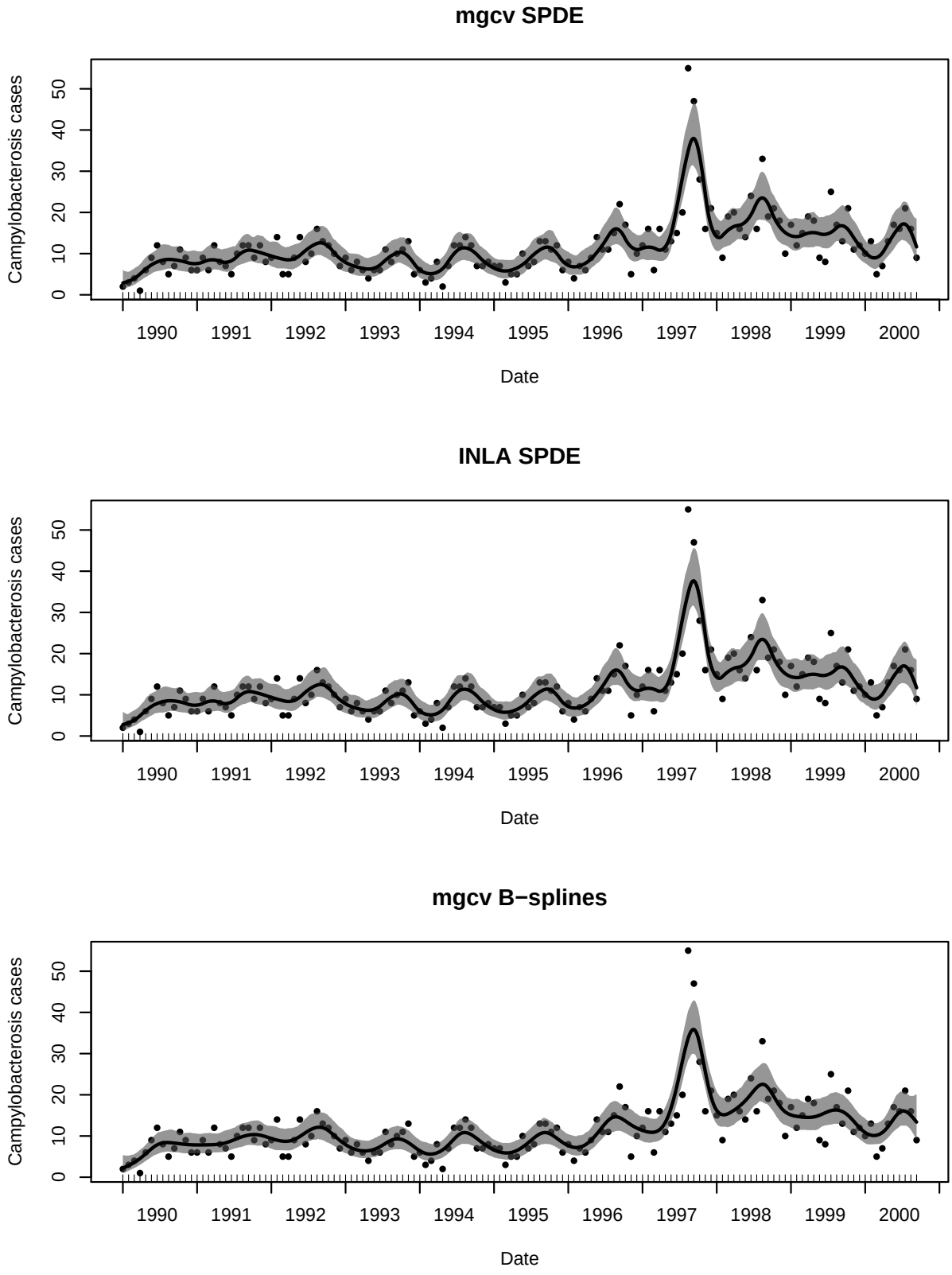


Figure 3: Campylobacterosis cases modelled using: a Matérn basis-penalty smoother fitted with `mgcv` (top), a Matérn SPDE fitted with `R-INLA` (middle), a B-spline basis-penalty smoother fitted using `mgcv` (bottom).

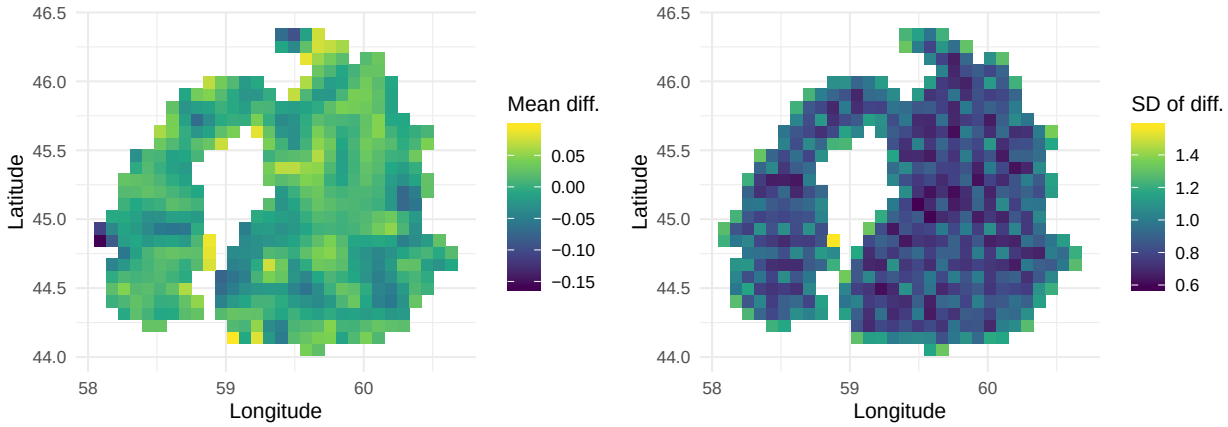


Figure 4: Chlorophyll in the Aral sea example. Left shows mean difference in predictions and right shows standard deviation of the difference in predictions between SPDE models fitted using `mgcv` and `R-INLA`.