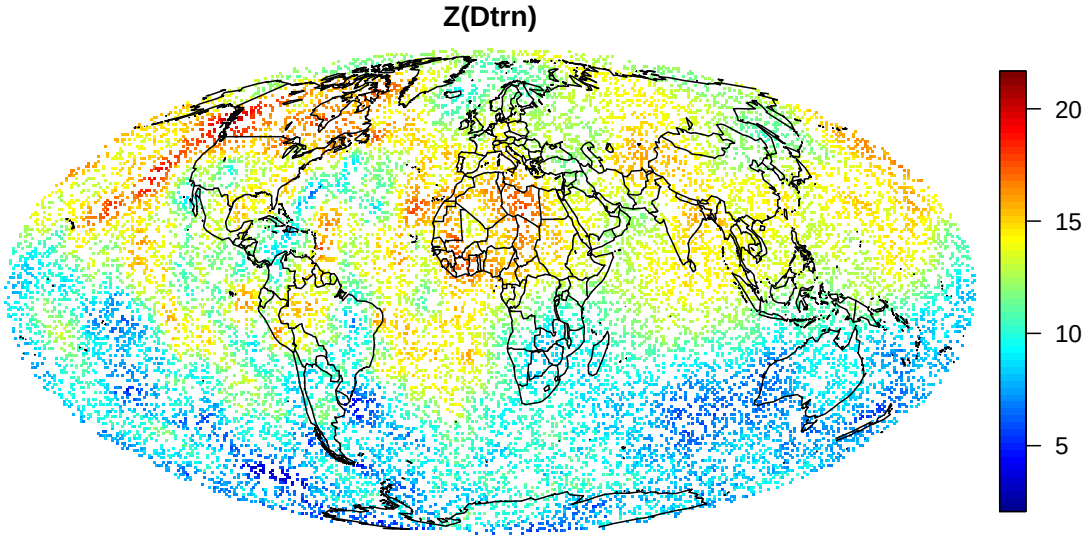
Numerical computation for statistics

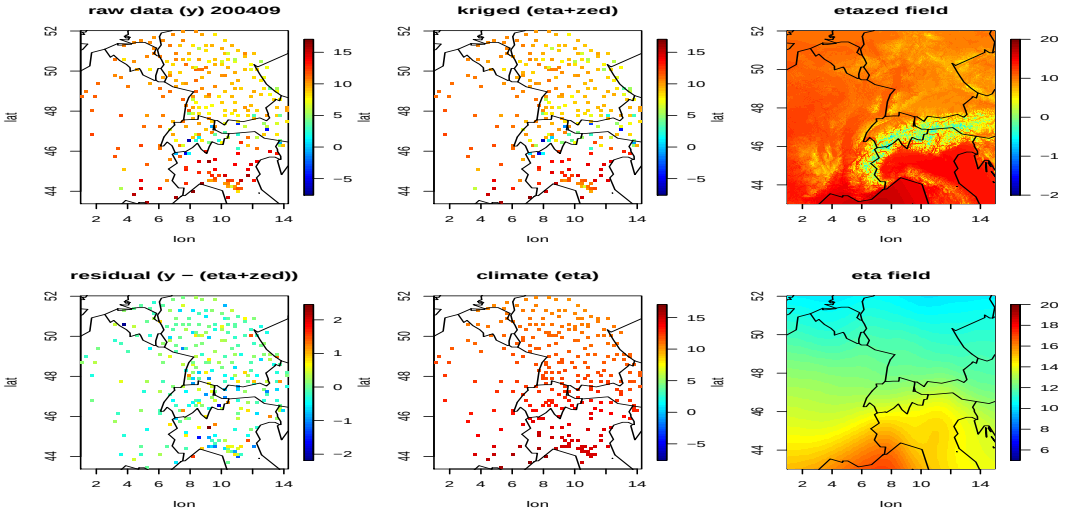
Finn Lindgren, University of Edinburgh, Scotland
http://www.maths.ed.ac.uk/~flindgre/cuso2019/
finn.lindgren@ed.ac.uk

CUSO, Villars-sur-Ollon, 1-4 September 2019

“Big” data



Sparse spatial coverage of temperature measurements



Regional observations: $\approx$ 20,000,000 from daily timeseries over 160 years

Overview

- Spatial statistics (but perhaps not like you're used to if you've seen it before)
- From models to numerics with the help of Markov in space
- MCMC-free Bayesian inference with direct numerical approximations
- Assessing numerical and approximate methods; principled method assessment
- Probabilistic model assessment with proper scoring rules
- Scaling it up; Likelihood and covariance matrix for a 10^{11}-dimensional vector? No thank you!
- Some R demonstrations (`INLA`, `inlabru`, `excursions`)

Spatio-temporal modelling framework

Spatial statistics framework

- Spatial domain D, or space-time domain $D \times \mathbb{T}$, $\mathbb{T} \subset \mathbb{R}$.
- Random field $u(\boldsymbol{s})$, $\boldsymbol{s} \in D$, or $u(\boldsymbol{s}, t)$, $(\boldsymbol{s}, t) \in D \times \mathbb{T}$.
- Observations y_i. In the simplest setting, $y_i = u(\boldsymbol{s}_i) + \epsilon_i$, but more generally $y_i \sim \mathsf{GLMM}$, with $u(\cdot)$ as a structured random effect.
- Needed: models capturing stochastic dependence on multiple scales
- Partial solution: Basis function expansions, with large scale functions and covariates to capture static and slow structures, and small scale functions for more local variability

Two basic model and method components

- Stochastic models for $u(\cdot)$.
- Computationally efficient (i.e. avoid MCMC whenever possible) inference methods for the posterior distribution of $u(\cdot)$ given data $\boldsymbol{y}$.

Covariance functions and stochastic PDEs

The Matérn covariance family on $\mathbb{R}^d$

$$\mathrm{Cov}(u(\mathbf{0}), u(\boldsymbol{s})) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} (\kappa \|\boldsymbol{s}\|)^\nu K_\nu(\kappa \|\boldsymbol{s}\|)$$

Scale $\kappa > 0$, smoothness $\nu > 0$, variance $\sigma^2 > 0$

Whittle (1954, 1963): Matérn as SPDE solution

Matérn fields are the stationary solutions to the SPDE

$$\left(\kappa^2 - \nabla \cdot \nabla\right)^{\alpha/2} u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \alpha = \nu + d/2$$

$\mathcal{W}(\cdot)$ white noise, $\nabla \cdot \nabla = \sum_{i=1}^d \frac{\partial^2}{\partial s_i^2}$, $\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)\kappa^{2\nu}(4\pi)^{d/2}}$

Gaussian random field (or Gaussian process)

A *Gaussian random field* $u : D \mapsto \mathbb{R}$ is defined via

$$\begin{aligned}
\mathsf{E}(u(\mathbf{s})) &= m(\mathbf{s}), \\
\mathsf{Cov}(u(\mathbf{s}), u(\mathbf{s}')) &= K(\mathbf{s}, \mathbf{s}'), \quad \text{(covariance kernel)} \\
\big[u(\mathbf{s}_i), i = 1, \ldots, n\big] &\sim \mathsf{N}(\boldsymbol{m} = \big[m(\mathbf{s}_i), i = 1, \ldots, n\big], \\
\boldsymbol{\Sigma} &= \big[K(\mathbf{s}_i, \mathbf{s}_j), i, j = 1, \ldots, n\big])
\end{aligned}$$

for all finite location sets $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, and $K(\cdot, \cdot)$ symmetric positive definite.

Generalised random field

A *generalised Gaussian random field* $u : D \mapsto \mathbb{R}$ is defined via a random measure, $\langle f, u\rangle_D = u^*(f) : H_{\mathcal{R}}(D) \mapsto \mathbb{R}$, $\mathcal{R}$ a covariance operator,

$$\mathsf{E}(\langle f, u\rangle_D) = \langle f, m\rangle_D = \int_D f(\mathbf{s}) m(\mathbf{s})\, \mathrm{d}\mathbf{s},$$

$$\mathsf{Cov}(\langle f, u\rangle_D, \langle g, u\rangle_D) = \langle f, \mathcal{R} g\rangle_D \equiv \iint_{D\times D} f(\mathbf{s}) K(\mathbf{s}, \mathbf{s}') g(\mathbf{s}')\, \mathrm{d}\mathbf{s}\, \mathrm{d}\mathbf{s}',$$

$$\langle f, u\rangle_D \sim \mathsf{N}(\langle f, m\rangle_D, \langle f, \mathcal{R} f\rangle_D)$$

for all $f, g \in H_{\mathcal{R}}(D) \equiv \{f : D \mapsto \mathbb{R};\ \langle f, \mathcal{R} f\rangle_D < \infty\}$.
This allows for singular covariance kernels $K(\cdot, \cdot)$.

White noise vs independent noise

Gaussian white noise on continuous domains

Standard Gaussian white noise $\mathcal{W}(\cdot)$ is a generalised random field, with

$$m(\mathbf{s}) = 0, \qquad K(\mathbf{s}, \mathbf{s}') = \delta_{\mathbf{s}}(\mathbf{s}'), \qquad \langle f, \mathcal{W} \rangle_D \sim \mathsf{N}(0, \langle f, f \rangle_D),$$

for all $f \in L_2(D)$. Since $\langle \delta_{\mathbf{s}}, \delta_{\mathbf{s}} \rangle_D = \infty$ for all $\mathbf{s} \in D$, $\mathcal{W}(\cdot)$ does not have pointwise meaning. We can only do calculus!

Independent Gaussian noise on continuous domains

Spatially independent Gaussian noise $w(\cdot)$ is a random field, with

$$m(\mathbf{s}) = 0, \qquad K(\mathbf{s}, \mathbf{s}') = \mathbf{1}_{\{\mathbf{s}=\mathbf{s}'\}}, \qquad w(\mathbf{s}) \sim \mathsf{N}(0, 1),$$

for all $\mathbf{s}, \mathbf{s}' \in D$. However, for every set $A \subset D$ with $|A|_{\mathsf{Leb(D)}} > 0$,

$$\mathsf{P}(\sup_{\mathbf{s} \in A} w(\mathbf{s}) = \infty) = \mathsf{P}(\inf_{\mathbf{s} \in A} w(\mathbf{s}) = -\infty) = 1,$$

and the generalised calculus is not applicable.

Spectral properties

Bochner's theorem on $\mathbb{R}^d$

A symmetric kernel $K(\mathbf{s}, \mathbf{s}')$, $\mathbf{s}, \mathbf{s}' \in \mathbb{R}^d$, is a positive (semi-)definite stationary covariance kernel if and only if there exists a non-negative spectral measure $S^*(\boldsymbol{\omega})$ such that

$$K(\mathbf{s}, \mathbf{s}') = \int_{\mathbb{R}^d} \exp(i(\mathbf{s}' - \mathbf{s}) \cdot \boldsymbol{\omega}) \, \mathrm{d}S^*(\boldsymbol{\omega})$$

If the measure has a density $S(\boldsymbol{\omega})$,

$$K(\mathbf{s}, \mathbf{s}') = \int_{\mathbb{R}^d} \exp(i(\mathbf{s}' - \mathbf{s}) \cdot \boldsymbol{\omega}) S(\boldsymbol{\omega}) \, \mathrm{d}\boldsymbol{\omega}$$

$$S(\boldsymbol{\omega}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \exp(-i\mathbf{s} \cdot \boldsymbol{\omega}) K(0, \mathbf{s}) \, \mathrm{d}\mathbf{s}$$

White noise on $\mathbb{R}^d$ has spectral density $S_{\mathcal{W}}(\boldsymbol{\omega}) = 1/(2\pi)^d$.

Spectral properties

Spectral representation

Let $Z^*(\boldsymbol{\omega})$ be a complex Gaussian random measure on $D = \mathbb{R}^d$ with independent increments and

$$\overline{\mathrm{d}Z^*(\boldsymbol{\omega})} = \mathrm{d}Z^*(-\boldsymbol{\omega}), \qquad \mathsf{E}\left[\mathrm{d}Z^*(\boldsymbol{\omega})\right] = 0, \qquad \mathsf{E}\left[\mathrm{d}Z^*(\boldsymbol{\omega})\overline{\mathrm{d}Z^*(\boldsymbol{\omega})}\right] = \mathrm{d}S^*(\boldsymbol{\omega}).$$

Then

$$u(\mathbf{s}) = \int_{\mathbb{R}^d} \exp(i\mathbf{s}\cdot\boldsymbol{\omega})\,\mathrm{d}Z^*(\boldsymbol{\omega})$$

is a stationary Gaussian random field with spectral measure $S^*(\boldsymbol{\omega})$.

Let $\widehat{f}(\boldsymbol{\omega}) = (\mathcal{F}f)(\boldsymbol{\omega}) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\exp(-i\mathbf{s}\cdot\boldsymbol{\omega})f(\mathbf{s})\,\mathrm{d}\mathbf{s}$.

Informally, $\widehat{u}(\boldsymbol{\omega})\,\mathrm{d}\boldsymbol{\omega} = \mathrm{d}Z^*(\boldsymbol{\omega})$, and the spectral density is $S_u(\boldsymbol{\omega}) = \mathsf{E}(|\widehat{u}(\boldsymbol{\omega})|^2)$.

Spectral properties

Spectra and linear differential operators

Differential operators can also be interpreted spectrally:

$\mathcal{L}f$	f	∇f	$-\nabla \cdot \nabla f$	$\mathcal{L}^{\alpha/2} f$
$\widehat{\mathcal{L}f} \equiv \mathcal{F}(\mathcal{L}f)$	$\widehat{f}$	$i\boldsymbol{\omega}\widehat{f}$	$\lVert\boldsymbol{\omega}\rVert^2\widehat{f}$	$\lvert\widehat{\mathcal{L}}\rvert^{\alpha/2}\widehat{f}$

The rightmost column is a *definition* of a fractional operator!

Exercise: Use the spectral field representation to derive the middle two results above.

Exercise: What would happen on a different manifold, such as the sphere? Hint: the harmonic functions in the Fourier transform are eigenfunctions of the Laplacian.

Covariance functions and stochastic PDEs

The Matérn covariance family on $\mathbb{R}^d$

$$\mathrm{Cov}(u(\mathbf{0}), u(\boldsymbol{s})) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} (\kappa \|\boldsymbol{s}\|)^\nu K_\nu(\kappa \|\boldsymbol{s}\|)$$

Scale $\kappa > 0$, smoothness $\nu > 0$, variance $\sigma^2 > 0$

Whittle (1954, 1963): Matérn as SPDE solution

Matérn fields are the stationary solutions to the SPDE

$$\left(\kappa^2 - \nabla \cdot \nabla\right)^{\alpha/2} u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \alpha = \nu + d/2$$

$\mathcal{W}(\cdot)$ white noise, $\nabla \cdot \nabla = \sum_{i=1}^d \frac{\partial^2}{\partial s_i^2}$, $\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)\kappa^{2\nu}(4\pi)^{d/2}}$

Spectral properties

For the Whittle-Matérn SPDE, informally,

$$
\begin{aligned}
(\kappa^2 - \nabla \cdot \nabla)^{\alpha/2} u(\mathbf{s}) &= \mathcal{W}(\mathbf{s}) \\
(\kappa^2 + \|\boldsymbol{\omega}\|^2)^{\alpha/2} \widehat{u}(\boldsymbol{\omega}) &= \widehat{\mathcal{W}}(\boldsymbol{\omega}) \\
\mathsf{E}(|(\kappa^2 + \|\boldsymbol{\omega}\|^2)^{\alpha/2} \widehat{u}(\boldsymbol{\omega})|^2) &= \mathsf{E}(|\widehat{\mathcal{W}}(\boldsymbol{\omega})|^2) \\
(\kappa^2 + \|\boldsymbol{\omega}\|^2)^{\alpha} S_u(\boldsymbol{\omega}) &= S_{\mathcal{W}}(\boldsymbol{\omega}) \\
S_u(\boldsymbol{\omega}) &= \frac{1}{(2\pi)^d (\kappa^2 + \|\boldsymbol{\omega}\|^2)^{\alpha}}
\end{aligned}
$$

Whittle (1954, 1963) showed that $K(\mathbf{s}, \mathbf{s}') = (\mathcal{F}^{-1} S_u(\cdot))(\mathbf{s}' - \mathbf{s})$ is equal to the Matérn covariance (up to a known scaling constant), with smoothness $\nu = \alpha - d/2$.

Simple heat equation

For space-time fields, we write $u(\mathbf{s},t)$, $(\mathbf{s},t) \in \mathbb{R}^d \times \mathbb{R}$, and $S_u(\boldsymbol{k},\omega)$, $(\boldsymbol{k},\omega) \in \mathbb{R}^d \times \mathbb{R}$.
We drive a heat equation with a noise process $\mathcal{E}$ that is white noise in time and Matérn noise in space, with parameters matching the heat operator:

$$\left\{\gamma\frac{\partial}{\partial t} + \kappa^2 - \nabla_s \cdot \nabla_s\right\} u(\mathbf{s}) = \mathcal{E}(\mathbf{s},t),$$
$$(\kappa^2 - \nabla_s \cdot \nabla_s)^{\alpha/2}\mathcal{E}(\mathbf{s},t) = \mathcal{W}(\mathbf{s},t).$$

The Fourier domain version is

$$\left\{i\gamma\omega + \kappa^2 + \|\boldsymbol{k}\|^2\right\}\widehat{u}(\boldsymbol{k},\omega) = \widehat{\mathcal{E}}(\boldsymbol{k},\omega),$$
$$(\kappa^2 + \|\boldsymbol{k}\|^2)^{\alpha/2}\widehat{\mathcal{E}}(\boldsymbol{k},\omega) = \widehat{\mathcal{W}}(\boldsymbol{k},\omega),$$

and

$$S_u(\boldsymbol{k},\omega) = \frac{1}{(2\pi)^{d+1}(\gamma^2\omega^2 + (\kappa^2 + \|\boldsymbol{k}\|^2)^2)(\kappa^2 + \|\boldsymbol{k}\|^2)^{\alpha}}$$

How differentiable are the realisations?

Simple heat equation (cont)

Using that, in the standardised Whittle-Matérn SPDE, the variance is

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)\kappa^{2\nu}(4\pi)^{d/2}}, \quad \nu = \alpha - d/2,$$

the marginal spatial spectrum for the heat model is

$$S_u(\boldsymbol{k}) = \int_{\mathbb{R}} S_u(\boldsymbol{k}, \omega)\, \mathrm{d}\omega = \frac{1}{4\pi\gamma} \frac{1}{(2\pi)^d(\kappa^2 + \|\boldsymbol{k}\|^2)^{\alpha+1}},$$

which is a scaled Whittle spectrum for a Matérn covariance with smoothness $\nu = \alpha + 1 - d/2$.

A generalised generalised case

If $\alpha = 0$, $d = 2$, then $\nu = 0$, which is just outside of the allowed range of the Matérn family. However, for every t, $u(\cdot, t)$ is a generalised random field with singular kernel $K(\mathbf{s}, \mathbf{s}') = \frac{1}{4\pi\gamma}\frac{1}{2\pi}K_0(\kappa\|\mathbf{s}' - \mathbf{s}\|)$.

Simple heat equation (cont)

To help understand the temporal properties, take the Fourier transform in only the spatial directions:

$$\left\{\gamma\frac{\partial}{\partial t}+\kappa^2+\|\boldsymbol{k}\|^2\right\}\widetilde{u}(\boldsymbol{k},t)=\frac{\widetilde{\mathcal{W}}(\boldsymbol{k},t)}{(\kappa^2+\|\boldsymbol{k}\|^2)^{\alpha/2}},$$

so for each spatial frequency $\boldsymbol{k}$, the temporal evolution of $\widetilde{u}(\boldsymbol{k},t)$ is an Ornstein-Uhlenbeck process with covariance

$$\frac{1}{4\pi\gamma(\kappa^2+\|\boldsymbol{k}\|^2)^{\alpha+1}}\exp\left(-|t|\frac{\kappa^2+\|\boldsymbol{k}\|^2}{\gamma}\right).$$

There is one more property we need to understand: Markov in space

First order Markov in time

Filtration σ-algebras:

$$a \in \mathcal{F}^\sigma_{(-\infty,t]} \equiv \sigma(u(s), s \le t), \quad b \in \mathcal{F}^\sigma_{[t,\infty)} \equiv \sigma(u(s), s \ge t)$$
$$\mathsf{P}(a \cap b \mid u(t)) = \mathsf{P}(a \mid u(t))\mathsf{P}(b \mid u(t))$$

Higher order Markov on spatial and spatio-temporal domains

Let $A, B, S \subset D$, such that S separates A and B.

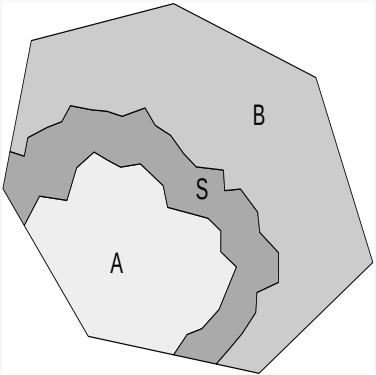
$$\mathcal{F}^\sigma_S \equiv \sigma(u(\mathbf{s}), \mathbf{s} \in S), \quad a \in \mathcal{F}^\sigma_A, \quad b \in \mathcal{F}^\sigma_B,$$
$$\mathsf{P}(a \cap b \mid \mathcal{F}^\sigma_S) = \mathsf{P}(a \mid \mathcal{F}^\sigma_S)\mathsf{P}(b \mid \mathcal{F}^\sigma_S)$$

Markov for generalised random fields

$$\mathcal{F}^\sigma_S \equiv \sigma(\langle f, u\rangle_S, f \in H_{\mathcal{R}}(S)), \quad a \in \mathcal{F}^\sigma_A, \quad b \in \mathcal{F}^\sigma_B,$$
$$\mathsf{P}(a \cap b \mid \mathcal{F}^\sigma_S) = \mathsf{P}(a \mid \mathcal{F}^\sigma_S)\mathsf{P}(b \mid \mathcal{F}^\sigma_S)$$

Markov in space

Markov properties



S is a separating set for A and B: $u(A) \perp u(B) \mid u(S)$

Solutions to

$$\left(\kappa^2 - \nabla \cdot \nabla\right)^{\alpha/2} u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s})$$

are Markov when α is an integer.

More generally, when the reciprocal of the spectral density is a polynomial, Rozanov, 1977

In graphs with no edges between A and B ($\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$):

$$\boldsymbol{Q}_{AB} = \boldsymbol{0}$$
$$\boldsymbol{Q}_{A|S,B} = \boldsymbol{Q}_{AA}$$
$$\boldsymbol{\mu}_{A|S,B} = \boldsymbol{\mu}_A - \boldsymbol{Q}_{AA}^{-1}\boldsymbol{Q}_{AS}(\boldsymbol{u}_S - \boldsymbol{\mu}_S)$$

Generally: Markov iff the precision *operator* $\mathcal{Q} = \mathcal{R}^{-1}$ is local.

Markov in space

The precision matrix block structure $\begin{bmatrix} \boldsymbol{Q}_{AA} & \boldsymbol{Q}_{AS} & \boldsymbol{0} \\ \boldsymbol{Q}_{SA} & \boldsymbol{Q}_{SS} & \boldsymbol{Q}_{SB} \\ \boldsymbol{0} & \boldsymbol{Q}_{BS} & \boldsymbol{Q}_{BB} \end{bmatrix}$ has important implications for practical computation (Cholesky, see later)

A partial history of Markov random fields

Rozanov (1977)

Generally: Markov iff the precision *operator* $\mathcal{Q} = \mathcal{R}^{-1}$ is local.

Stationary case:

$$u(\mathbf{s}) \text{ is stationary Markov} \iff S_u(\boldsymbol{k}) \propto P(\boldsymbol{k})^{-1}$$

where $P(\boldsymbol{k}) \geq 0$ is a symmetric polynomial

Matérn/Whittle is Markov for $\alpha = 1, 2, 3 \ldots$: $S_u(\boldsymbol{k}) \propto (\kappa^2 + \|\boldsymbol{k}\|^2)^{-\alpha}$

GMRF	Covariance on $\mathbb{R}^2$	
SAR(1), CAR(2)	$\propto \kappa\|\boldsymbol{u}\| \, K_1(\kappa\|\boldsymbol{s} - \boldsymbol{s}'\|)$	Whittle (1954)
CAR(1)	$\frac{1}{2\pi} K_0(\kappa\|\boldsymbol{s} - \boldsymbol{s}'\|)$	Besag (1981)
ICAR(1)	$-\frac{1}{2\pi} \log(\|\boldsymbol{s} - \boldsymbol{s}'\|)$	Besag & Mondal (2005)

On lattices, classical CAR $\rightarrow$ Matérn models (limits of).

Hilbert space approximation ("The SPDE approach" from Lindgren et al, 2011)

Can extend to (non-)stationary SPDE models on irregular triangulations.

From continuous to discrete

We want to construct finite dimensional approximations to the distribution of $u(\cdot)$, where

$$\left[\left\langle f_i, (\kappa^2 - \nabla \cdot \nabla)^{\alpha/2} u(\cdot)\right\rangle_D, i = 1, \ldots, m\right] \stackrel{d}{=} \left[\left\langle f_i, \mathcal{W}(\cdot)\right\rangle_D, i = 1, \ldots, m\right]$$

for all finite collections of test functions $f_i \in H_{\mathcal{R}_\mathcal{W}}(D)$.

A finite basis expansion

$$u(\mathbf{s}) = \sum_{j=1}^{n} \psi_j(\mathbf{s}) u_j$$

can only hope to achieve this for a subspace of size n.

Two main approaches:

- Galerkin: $\{f_i = \psi_i,\ i = 1, \ldots, n\}$
- Least squares: $\{f_i = (\kappa^2 - \nabla \cdot \nabla)^{\alpha/2} \psi_i,\ i = 1, \ldots, n\}$

We use least squares for $\alpha = 1$, Galerkin for $\alpha = 2$, and a recursion for $\alpha \geq 3$.

Stochastic Green's first identity

On any sufficiently smooth manifold domain D,

$$\langle f, -\nabla \cdot \nabla g \rangle_D = \langle \nabla f, \nabla g \rangle_D - \langle f, \partial_n g \rangle_{\partial D}$$

holds, even if either ∇f or $-\nabla \cdot \nabla g$ are as generalised as white noise.

For now, we'll impose deterministic Neumann boundary conditions, informally $\partial_n u(\mathbf{s}) = 0$ for all $\mathbf{s} \in \partial D$. For $\alpha = 2$ and Galerkin,

$$\begin{aligned}\left\langle \psi_i, (\kappa^2 - \nabla \cdot \nabla) \sum_j \psi_j u_j \right\rangle_D &= \sum_j \left\{ \kappa^2 \langle \psi_i, \psi_j \rangle_D + \langle \nabla \psi_i, \nabla \psi_j \rangle_D \right\} u_j \\ &= (\kappa^2 \boldsymbol{C} + \boldsymbol{G}) \boldsymbol{u}\end{aligned}$$

The covariance for the RHS of the SPDE is

$$\left[\mathsf{Cov}(\langle \psi_i, \mathcal{W} \rangle_D, \langle \psi_j, \mathcal{W} \rangle_D \right] = \left[\langle \psi_i, \psi_j \rangle_D \right] = \boldsymbol{C}$$

by the definition of $\mathcal{W}$.

We seek $\boldsymbol{u} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ such that $\mathsf{Var}\{(\kappa^2\boldsymbol{C} + \boldsymbol{G})\boldsymbol{u}\} = \boldsymbol{C}$:

$$(\kappa^2\boldsymbol{C} + \boldsymbol{G})\boldsymbol{\Sigma}(\kappa^2\boldsymbol{C} + \boldsymbol{G}) = \boldsymbol{C}$$
$$\boldsymbol{\Sigma} = (\kappa^2\boldsymbol{C} + \boldsymbol{G})^{-1}\boldsymbol{C}(\kappa^2\boldsymbol{C} + \boldsymbol{G})^{-1}$$

If ψ_i are piecewise linear on a triangulation of D, then $\boldsymbol{C}$ and $\boldsymbol{G}$ are both very sparse, and in addition, $\boldsymbol{C} = \mathrm{diag}(\langle \psi_i, 1 \rangle_D)$ is a valid approximation. Then, the *precision* matrix is also sparse,

$$\boldsymbol{Q} = (\kappa^2\boldsymbol{C} + \boldsymbol{G})\boldsymbol{C}^{-1}(\kappa^2\boldsymbol{C} + \boldsymbol{G})$$

and $\boldsymbol{u}$ is Markov on the adjacency graph given by the non-zero structure of $\boldsymbol{Q}$.

Least squares and Galerkin recursion gives precisions for all $\alpha = 1, 2, \ldots$:

- $\boldsymbol{Q}_1 = (\kappa^2\boldsymbol{C} + \boldsymbol{G})$
- $\boldsymbol{Q}_2 = (\kappa^2\boldsymbol{C} + \boldsymbol{G})\boldsymbol{C}^{-1}(\kappa^2\boldsymbol{C} + \boldsymbol{G}) = \kappa^4\boldsymbol{C} + 2\kappa^2\boldsymbol{G} + \boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}$
- $\boldsymbol{Q}_\alpha = (\kappa^2\boldsymbol{C} + \boldsymbol{G})\boldsymbol{C}^{-1}\boldsymbol{Q}_{\alpha-2}\boldsymbol{C}^{-1}(\kappa^2\boldsymbol{C} + \boldsymbol{G})$
- Any $\alpha \geq 0$: $\boldsymbol{Q}_\alpha = \boldsymbol{C}^{1/2}\left\{\boldsymbol{C}^{-1/2}(\kappa^2\boldsymbol{C} + \boldsymbol{G})\boldsymbol{C}^{-1/2}\right\}^\alpha \boldsymbol{C}^{1/2}$
  (non-sparse for non-integer α)

Basis function representations for Gaussian Matérn fields

Basis definitions

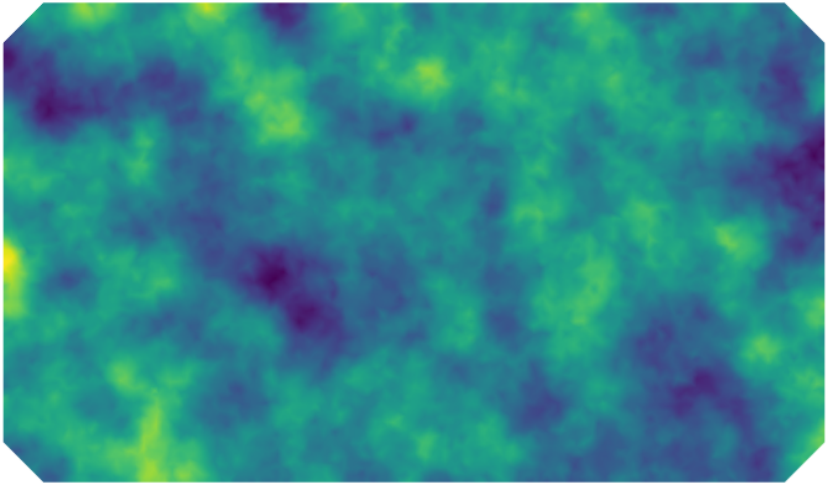
	Finite basis set ($k = 1, \ldots, n$)
Karhunen-Loève	$(\kappa^2 - \nabla \cdot \nabla)^{-\alpha} e_{\kappa,k}(\boldsymbol{s}) = \lambda_{\kappa,k} e_{\kappa,k}(\boldsymbol{s})$
Fourier	$-\nabla \cdot \nabla e_k(\boldsymbol{s}) = \lambda_k e_k(\boldsymbol{s})$
Convolution	$(\kappa^2 - \nabla \cdot \nabla)^{\alpha/2} g_\kappa(\boldsymbol{s}) = \delta(\boldsymbol{s})$
General	$\psi_k(\boldsymbol{s})$

Field representations

	Field $u(\boldsymbol{s})$	Weights
Karhunen-Loève	$\propto \sum_k e_{\kappa,k}(\boldsymbol{s}) z_k$	$z_k \sim \mathsf{N}(0, \lambda_{\kappa,k})$
Fourier	$\propto \sum_k e_k(\boldsymbol{s}) z_k$	$z_k \sim \mathsf{N}(0, (\kappa^2 + \lambda_k)^{-\alpha})$
Convolution	$\propto \sum_k g_\kappa(\boldsymbol{s} - \boldsymbol{s}_k) z_k$	$z_k \sim \mathsf{N}(0, \lvert\mathsf{cell}_k\rvert)$
General	$\propto \sum_k \psi_k(\boldsymbol{s}) u_k$	$\boldsymbol{u} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{Q}_\kappa^{-1})$

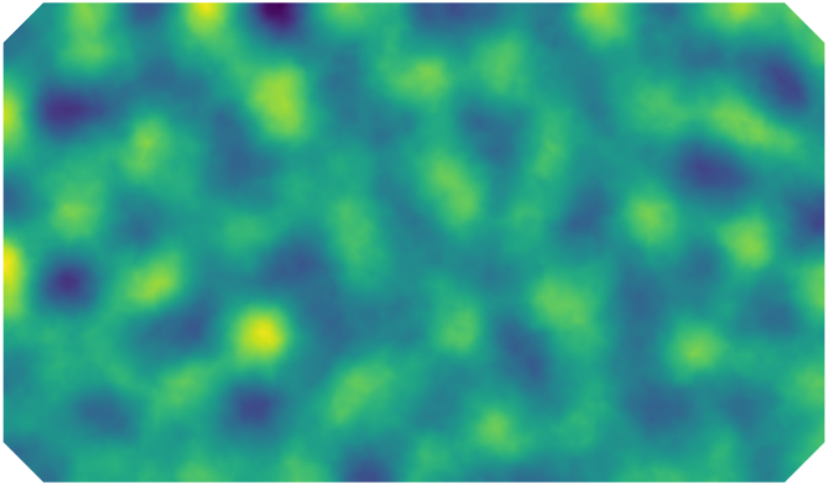
Note: Harmonic basis functions (as in the Fourier approach) give a diagonal $\boldsymbol{Q}_\kappa$, but lead to dense posterior precision matrices.

SPDE/GMRF realisations and non-stationary models



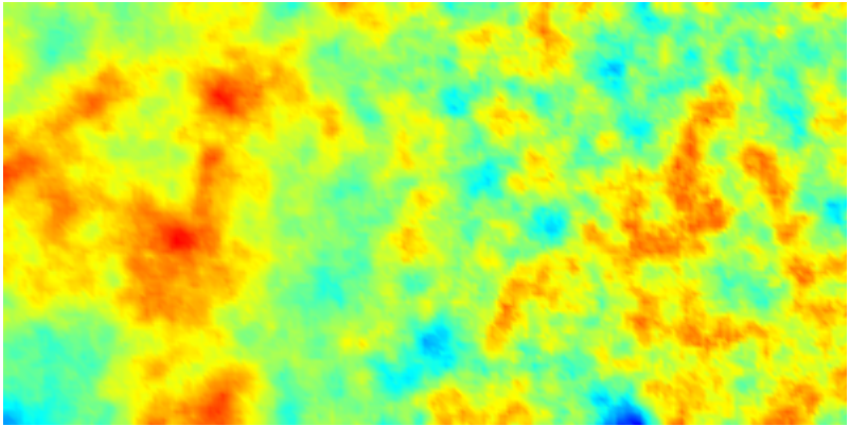
$$(\kappa^2 - \nabla \cdot \nabla) u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in D$$

SPDE/GMRF realisations and non-stationary models



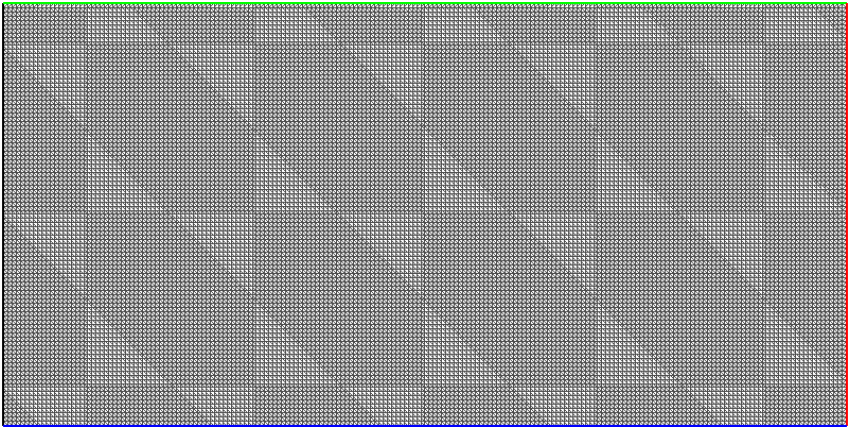
$(\kappa^2 \exp(i\theta) - \nabla \cdot \nabla)u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s})$, $\boldsymbol{s} \in D$, $\mathsf{Re}(u)$ independent of $\mathsf{Im}(u)$

Link to Sampson&Guttorp (1992) deformation non-stationarity



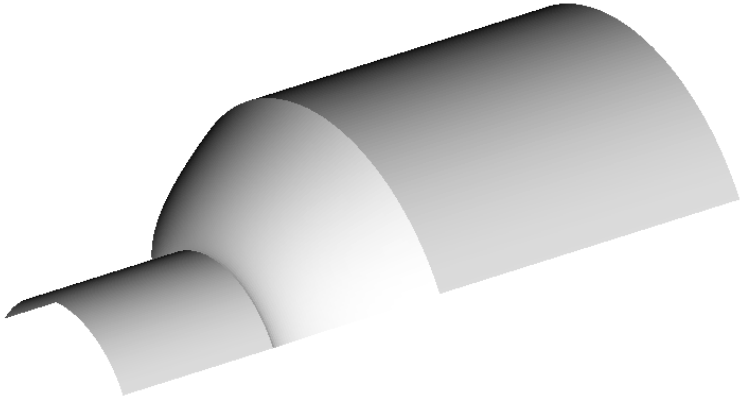
$$(\kappa(\boldsymbol{s})^2 - \nabla \cdot \nabla)u(\boldsymbol{s}) = \kappa(\boldsymbol{s})\mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \Omega$$

Link to Sampson&Guttorp (1992) deformation non-stationarity



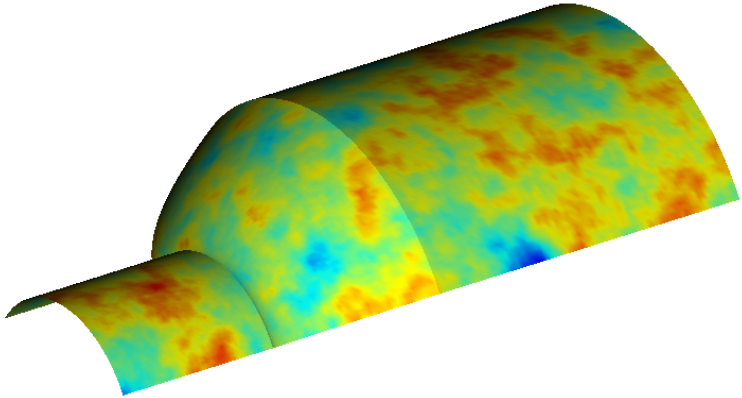
$$(\kappa(\boldsymbol{s})^2 - \nabla \cdot \nabla)u(\boldsymbol{s}) = \kappa(\boldsymbol{s})\mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \Omega$$

Link to Sampson&Guttorp (1992) deformation non-stationarity



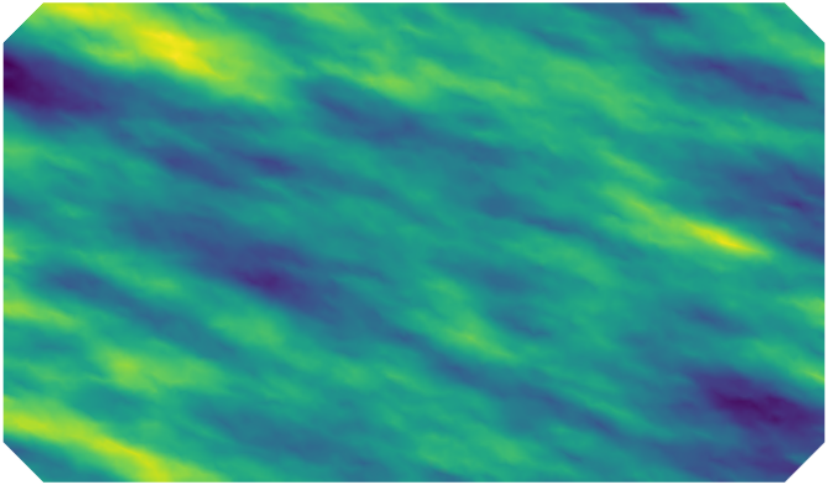
$$(\widetilde{\kappa}^2 - \nabla \cdot \nabla) u(\widetilde{\boldsymbol{s}}) = \widetilde{\kappa}\widetilde{\mathcal{W}}(\widetilde{\boldsymbol{s}}), \quad \widetilde{\boldsymbol{s}} \in \widetilde{\Omega}$$

Link to Sampson&Guttorp (1992) deformation non-stationarity



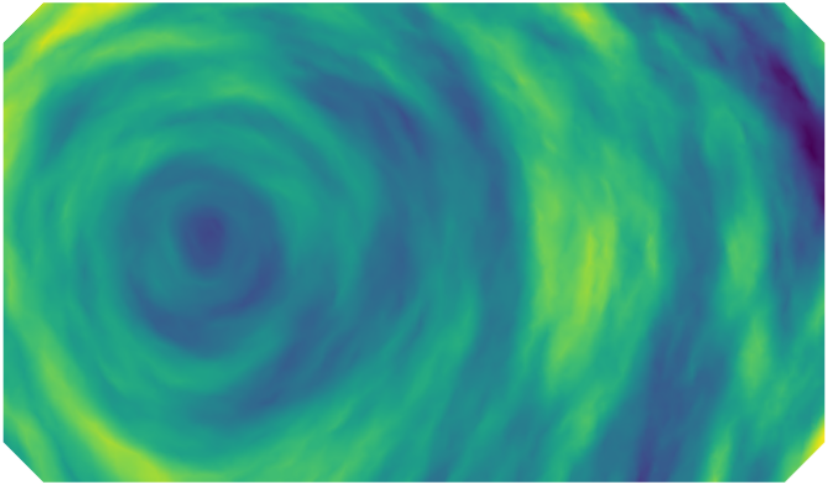
$$(\widetilde{\kappa}^2 - \nabla \cdot \nabla) u(\widetilde{\boldsymbol{s}}) = \widetilde{\kappa}\widetilde{\mathcal{W}}(\widetilde{\boldsymbol{s}}), \quad \widetilde{\boldsymbol{s}} \in \widetilde{\Omega}$$

SPDE/GMRF realisations and non-stationary models



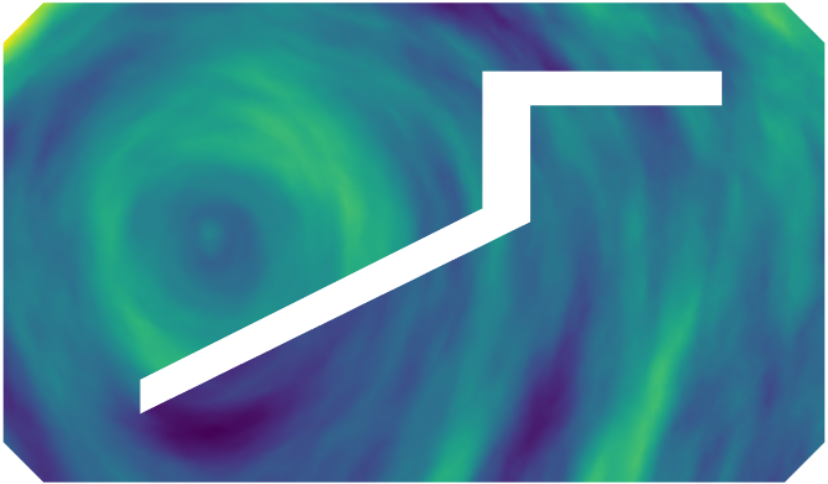
$$(\kappa^2 - \nabla \cdot \boldsymbol{H} \nabla) u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in D$$

SPDE/GMRF realisations and non-stationary models



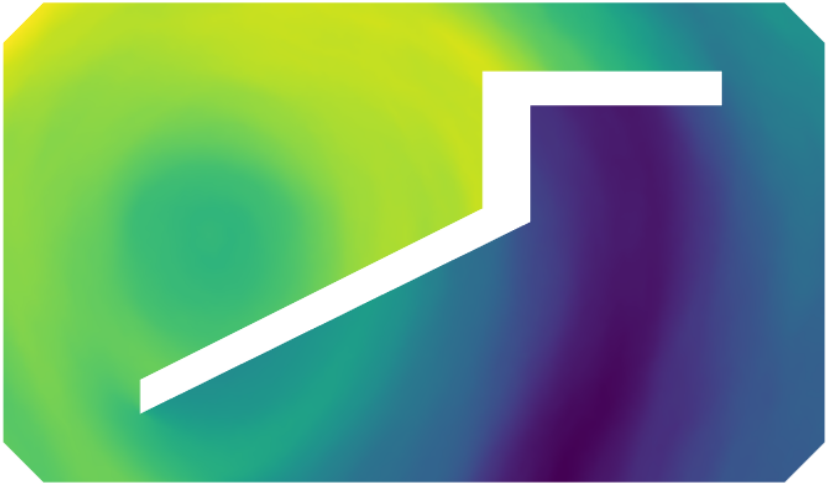
$$(\kappa^2 - \nabla \cdot \boldsymbol{H}(\boldsymbol{s})\nabla)u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in D$$

SPDE/GMRF realisations and non-stationary models



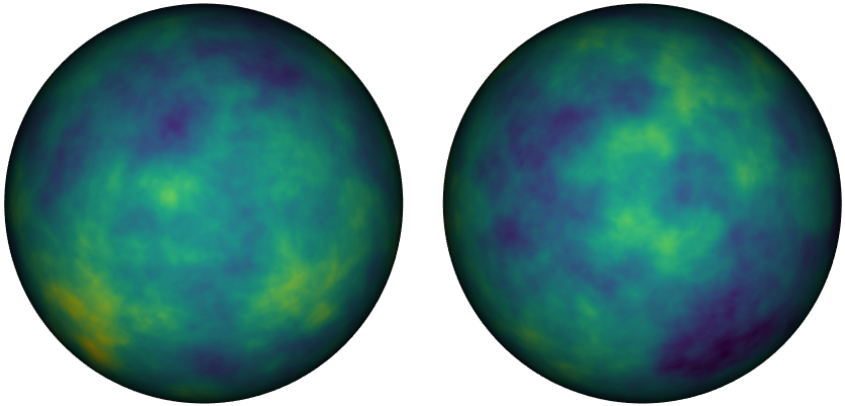
$$(\kappa^2 - \nabla \cdot \nabla) u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in D$$

SPDE/GMRF realisations and non-stationary models



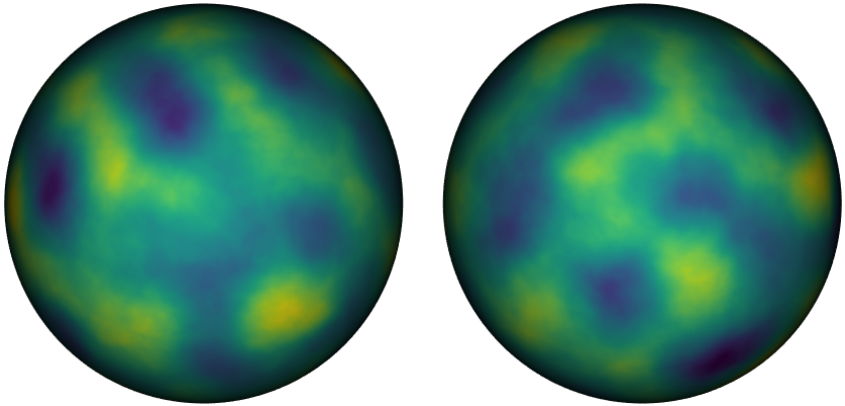
$$(\kappa^2 - \nabla \cdot \nabla) u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in D$$

SPDE/GMRF realisations and non-stationary models



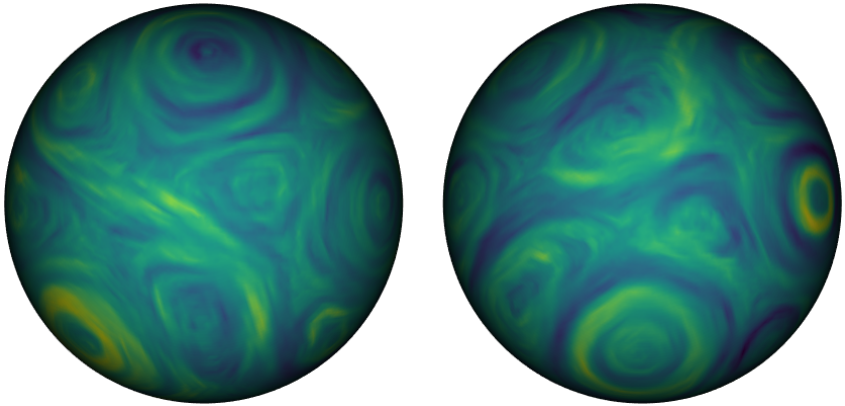
$$(\kappa^2 - \nabla \cdot \nabla) u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in D = \mathbb{S}^2$$

SPDE/GMRF realisations and non-stationary models



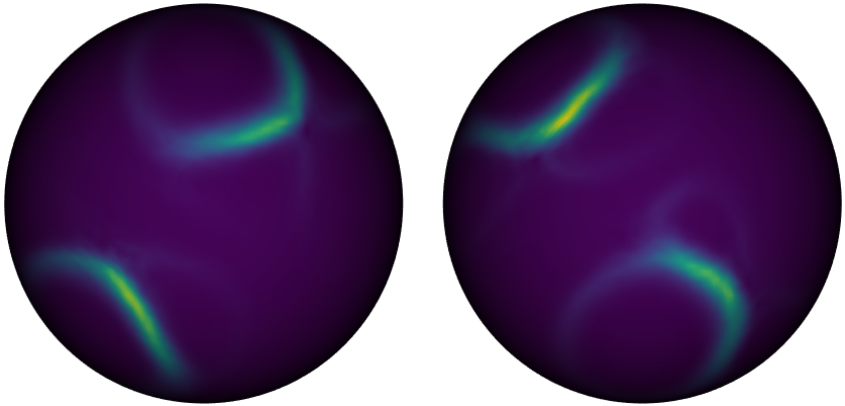
$$(\kappa^2 \exp(i\theta) - \nabla \cdot \nabla) u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in D = \mathbb{S}^2$$

Markov does *not* mean that dependence is only local



$$(\kappa(\boldsymbol{s})^2 - \nabla \cdot \boldsymbol{H}(\boldsymbol{s})\nabla)u(\boldsymbol{s}) = \kappa(\boldsymbol{s})\mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \Omega$$

Covariances for four reference points



Hierarchical models

Continuous Markovian spatial models (Lindgren et al, 2011)

Local basis: $u(\boldsymbol{s}) = \sum_k \psi_k(\boldsymbol{s}) u_k, \quad$ (compact, piecewise linear)

Basis weights: $\boldsymbol{u} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1}), \quad$ sparse $\boldsymbol{Q}$ based on an SPDE

Special case: $(\kappa^2 - \nabla \cdot \nabla) u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \Omega$

Precision: $\boldsymbol{Q} = \kappa^4 \boldsymbol{C} + 2\kappa^2 \boldsymbol{G} + \boldsymbol{G}_2 \qquad (\kappa^4 + 2\kappa^2 |\boldsymbol{\omega}|^2 + |\boldsymbol{\omega}|^4)$

Conditional distribution in a jointly Gaussian model

$$\boldsymbol{u} \sim \mathsf{N}(\boldsymbol{\mu}_u, \boldsymbol{Q}_u^{-1}), \quad \boldsymbol{y}|\boldsymbol{u} \sim \mathsf{N}(\boldsymbol{A}\boldsymbol{u}, \boldsymbol{Q}_{y|u}^{-1}) \qquad (A_{ij} = \psi_j(\boldsymbol{s}_i))$$

$$\boldsymbol{u}|\boldsymbol{y} \sim \mathsf{N}(\boldsymbol{\mu}_{u|y}, \boldsymbol{Q}_{u|y}^{-1})$$

$$\boldsymbol{Q}_{u|y} = \boldsymbol{Q}_u + \boldsymbol{A}^T \boldsymbol{Q}_{y|u} \boldsymbol{A} \quad (\sim\text{"Sparse iff } \psi_k \text{ have compact support"})$$

$$\boldsymbol{\mu}_{u|y} = \boldsymbol{\mu}_u + \boldsymbol{Q}_{u|y}^{-1} \boldsymbol{A}^T \boldsymbol{Q}_{y|u} (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}_u)$$

The computational GMRF work-horse

Cholesky decomposition (Cholesky, 1924)

$$\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^\top, \quad \boldsymbol{L} \text{ lower triangular } (\sim \mathcal{O}(n^{(d+1)/2}) \text{ for } d = 1, 2, 3)$$

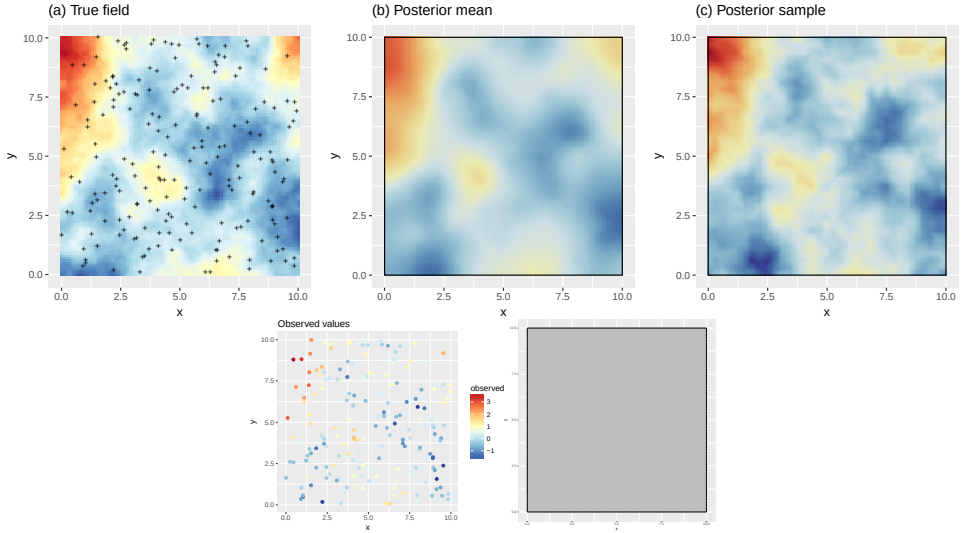
$$\boldsymbol{Q}^{-1}\boldsymbol{x} = \boldsymbol{L}^{-\top}\boldsymbol{L}^{-1}\boldsymbol{x}, \quad \text{via forward/backward substitution}$$

$$\log\det \boldsymbol{Q} = 2\log\det \boldsymbol{L} = 2\sum_i \log L_{ii}$$

André-Louis Cholesky (1875–1918)

"He invented, for the solution of the condition equations in the method of least squares, a very ingenious computational procedure which immediately proved extremely useful, and which most assuredly would have great benefits for all geodesists, if it were published some day." (Euology by Commandant Benoit, 1922)

Example: 2D georeferenced data



How to choose a triangulation mesh?

- SPDE solutions with Neumann boundary conditions are not stationary processes; there is a boundary effect on the covariance structure; visible as inflated variance (factor 2 for straight boundaries)
- Easy workaround: extend the domain boundary
- Small triangles lead to good continuous function approximation properties
- Small triangles lead to expensive calculations
- Resolve the tradeoff by choosing the triangles to be *small enough* in relation to the correlation length. Need intuition!
- Exercise: Given $\mathsf{E}(u_0) = \mathsf{E}(u_1) = 0$, $\mathsf{Var}(u_0) = \sigma_0^2$, $\mathsf{Var}(u_1) = \sigma_1^2$, and $\mathsf{Cov}(u_0, u_1) = \rho\sigma_0\sigma_1$, what is the variance of the linear interpolation $(1-z)u_0 + zu_1$, $z \in [0,1]$?
- When the triangle edge lengths decrease, the "ρ" values increase and the continuous/discrete model discrepancy decreases. This can be visualised:
  The interactive tool `INLA::meshbuilder()` can help build intuition

A multiscale model example

- A temporally slow, stochastic heat equation (non-separable)

$$\frac{\partial}{\partial t} z(\boldsymbol{s}, t) + \gamma_z (1 - \gamma_{\mathcal{E}} \nabla \cdot \nabla) z(\boldsymbol{s}, t) = \mathcal{E}(\boldsymbol{s}, t)$$
$$(1 - \gamma_{\mathcal{E}} \nabla \cdot \nabla)^{1/2} \mathcal{E}(\boldsymbol{s}, t) = \mathcal{W}_{\mathcal{E}}(\boldsymbol{s}, t)$$

- A temporally quick, spatially non-stationary SPDE/GMRF (separable)
  $\left(\frac{\partial}{\partial t} + \gamma_t\right) \left(\kappa(\boldsymbol{s})^2 - \nabla \cdot \nabla\right) \left(\tau(\boldsymbol{s}) a(\boldsymbol{s}, t)\right) = \mathcal{W}_a(\boldsymbol{s}, t)$
- Measurements
  $y_i = a(\boldsymbol{s}_i, t_i) + z(\boldsymbol{s}_i, t_i) + \epsilon_i$, discretised into
  $\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{a} + (\boldsymbol{B} \otimes \boldsymbol{I})\boldsymbol{z}) + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{Q}_\epsilon^{-1})$
  where $\boldsymbol{B}$ maps from long-term basis functions to short-term, and $\boldsymbol{A}$ maps from short-term basis functions to the observations.

The posterior precision can be formulated for $(a + z, z)|y$:

$$\boldsymbol{Q}_{(a+z,z)|y} = \begin{bmatrix} \boldsymbol{Q}_t \otimes \boldsymbol{Q}_a + \boldsymbol{A}^\top \boldsymbol{Q}_\epsilon \boldsymbol{A} & -\boldsymbol{Q}_t \boldsymbol{B} \otimes \boldsymbol{Q}_a \\ -\boldsymbol{B}^\top \boldsymbol{Q}_t \otimes \boldsymbol{Q}_a & \boldsymbol{Q}_z + \boldsymbol{B}^\top \boldsymbol{Q}_t \boldsymbol{B} \otimes \boldsymbol{Q}_a \end{bmatrix}$$

Locally isotropic non-stationary precision construction

Finite element construction of basis weight precision

Non-stationary SPDE:

$$(\kappa(\boldsymbol{s})^2 - \nabla \cdot \nabla)\left(\tau(\boldsymbol{s})u(\boldsymbol{s})\right) = \mathcal{W}(\boldsymbol{s})$$

The SPDE parameters are constructed via spatial covariates:

$$\log \tau(\boldsymbol{s}) = b_0^\tau(\boldsymbol{s}) + \sum_{j=1}^{p} b_j^\tau(\boldsymbol{s})\theta_j, \quad \log \kappa(\boldsymbol{s}) = b_0^\kappa(\boldsymbol{s}) + \sum_{j=1}^{p} b_j^\kappa(\boldsymbol{s})\theta_j$$

Finite element calculations give

$$\boldsymbol{T} = \operatorname{diag}(\tau(\boldsymbol{s}_i)), \quad \boldsymbol{K} = \operatorname{diag}(\kappa(\boldsymbol{s}_i))$$

$$C_{ii} = \int \psi_i(\boldsymbol{s})\, d\boldsymbol{s}, \quad G_{ij} = \int \nabla\psi_i(\boldsymbol{s}) \cdot \nabla\psi_j(\boldsymbol{s})\, d\boldsymbol{s}$$

$$\boldsymbol{Q} = \boldsymbol{T}\left(\boldsymbol{K}^2\boldsymbol{C}\boldsymbol{K}^2 + \boldsymbol{K}^2\boldsymbol{G} + \boldsymbol{G}\boldsymbol{K}^2 + \boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}\right)\boldsymbol{T}$$

Combining this with an AR(1) discretisation of the temporal operator, we get $\boldsymbol{Q}_t \otimes \boldsymbol{Q}_a$.

GMRF precision for stochastic heat equation

$$
\begin{aligned}
\boldsymbol{Q}_z &= \boldsymbol{M}_2^{(t)} \otimes \boldsymbol{M}_0^{(s)} + \boldsymbol{M}_1^{(t)} \otimes \boldsymbol{M}_1^{(s)} + \boldsymbol{M}_0^{(t)} \otimes \boldsymbol{M}_2^{(s)} \\
\boldsymbol{M}_0^{(s)} &= \boldsymbol{C} + \gamma_{\mathcal{E}} \boldsymbol{G} \\
\boldsymbol{M}_1^{(s)} &= \gamma_z (\boldsymbol{C} + \gamma_{\mathcal{E}} \boldsymbol{G}) \boldsymbol{C}^{-1} (\boldsymbol{C} + \gamma_{\mathcal{E}} \boldsymbol{G}) \\
\boldsymbol{M}_2^{(s)} &= \gamma_z^2 (\boldsymbol{C} + \gamma_{\mathcal{E}} \boldsymbol{G}) \boldsymbol{C}^{-1} (\boldsymbol{C} + \gamma_{\mathcal{E}} \boldsymbol{G}) \boldsymbol{C}^{-1} (\boldsymbol{C} + \gamma_{\mathcal{E}} \boldsymbol{G})
\end{aligned}
$$

The precision structure can be used to formulate sampling as

$$
\boldsymbol{Q}_z \boldsymbol{z} = \widetilde{\boldsymbol{L}}_z \boldsymbol{w}, \quad \boldsymbol{w} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I})
$$

where $\widetilde{\boldsymbol{L}}_z$ is a pseudo Cholesky factor,

$$
\begin{aligned}
\widetilde{\boldsymbol{L}}_z = \Big[\Big[& \boldsymbol{L}_2^{(t)} \otimes \boldsymbol{L}_{\boldsymbol{C}}, \quad \boldsymbol{L}_1^{(t)} \otimes \boldsymbol{L}_{\boldsymbol{G}}, \quad \boldsymbol{L}_0^{(t)} \otimes \boldsymbol{G} \boldsymbol{L}_{\boldsymbol{C}}^{-\top} \Big], \\
& \gamma_{\mathcal{E}}^{1/2} \Big[\boldsymbol{L}_2^{(t)} \otimes \boldsymbol{L}_{\boldsymbol{G}}, \quad \boldsymbol{L}_1^{(t)} \otimes \boldsymbol{G} \boldsymbol{L}_{\boldsymbol{C}}^{-\top}, \quad \boldsymbol{L}_0^{(t)} \otimes \boldsymbol{G} \boldsymbol{C}^{-1} \boldsymbol{L}_{\boldsymbol{G}} \Big] \Big]
\end{aligned}
$$

Posterior calculations

Write $x = (a + z, z)$ for the full latent field.

$$\boldsymbol{Q}_{x|y} = \begin{bmatrix} \boldsymbol{Q}_t \otimes \boldsymbol{Q}_a + \boldsymbol{A}^\top \boldsymbol{Q}_\epsilon \boldsymbol{A} & -\boldsymbol{Q}_t \boldsymbol{B} \otimes \boldsymbol{Q}_a \\ -\boldsymbol{B}^\top \boldsymbol{Q}_t \otimes \boldsymbol{Q}_a & \boldsymbol{Q}_z + \boldsymbol{B}^\top \boldsymbol{Q}_t \boldsymbol{B} \otimes \boldsymbol{Q}_a \end{bmatrix}$$

can be pseudo-Cholesky-factorised:

$$\boldsymbol{Q}_{x|y} = \widetilde{\boldsymbol{L}}_{x|y} \widetilde{\boldsymbol{L}}_{x|y}^\top, \qquad \widetilde{\boldsymbol{L}}_{x|y} = \begin{bmatrix} \boldsymbol{L}_t \otimes \boldsymbol{L}_a & \boldsymbol{0} & \boldsymbol{A}^\top \boldsymbol{L}_\epsilon \\ -\boldsymbol{B}^\top \boldsymbol{L}_t \otimes \boldsymbol{L}_a & \widetilde{\boldsymbol{L}}_z & \boldsymbol{0} \end{bmatrix}$$

Posterior expectation, samples, and marginal variances (with $\widetilde{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{0} \end{bmatrix}$):

$$\begin{aligned} \boldsymbol{Q}_{x|y}(\boldsymbol{\mu}_{x|y} - \boldsymbol{\mu}_x) &= \widetilde{\boldsymbol{A}}^\top \boldsymbol{Q}_\epsilon (\boldsymbol{y} - \widetilde{\boldsymbol{A}} \boldsymbol{\mu}_x), \\ \boldsymbol{Q}_{x|y}(\boldsymbol{x} - \boldsymbol{\mu}_{x|y}) &= \widetilde{\boldsymbol{L}}_{x|y} \boldsymbol{w}, \quad \boldsymbol{w} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}), \quad \text{or} \\ \boldsymbol{Q}_{x|y}(\boldsymbol{x} - \boldsymbol{\mu}_x) &= \widetilde{\boldsymbol{A}}^\top \boldsymbol{Q}_\epsilon (\boldsymbol{y} - \widetilde{\boldsymbol{A}} \boldsymbol{\mu}_x) + \widetilde{\boldsymbol{L}}_{x|y} \boldsymbol{w}, \quad \boldsymbol{w} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}), \\ \mathsf{Var}(x_i|\boldsymbol{y}) &= \left(\boldsymbol{Q}_{x|y}^{-1} \right)_{ii} \quad \text{(Ouch! Don't do this!! Use Takahashi!!!)} \end{aligned}$$

Part 2: Fast Bayesian inference & method and model assessment

Laplace approximations for non-Gaussian observations

Quadratic posterior log-likelihood approximation

$$
\begin{aligned}
p(\boldsymbol{u} \mid \boldsymbol{\theta}) &\sim \mathsf{N}(\boldsymbol{\mu}_u, \boldsymbol{Q}_u^{-1}), \quad \boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\theta} \sim p(\boldsymbol{y} \mid \boldsymbol{u}) \\
p_G(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta}) &\sim \mathsf{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{Q}}^{-1}) \\
\mathbf{0} &= \nabla_{\boldsymbol{u}} \left\{ \ln p(\boldsymbol{u} \mid \boldsymbol{\theta}) + \ln p(\boldsymbol{y} \mid \boldsymbol{u}) \right\} \big|_{\boldsymbol{u}=\widetilde{\boldsymbol{\mu}}} \\
\widetilde{\boldsymbol{Q}} &= \boldsymbol{Q}_u - \nabla_{\boldsymbol{u}}^2 \ln p(\boldsymbol{y} \mid \boldsymbol{u}) \big|_{\boldsymbol{u}=\widetilde{\boldsymbol{\mu}}}
\end{aligned}
$$

Direct Bayesian inference with INLA (`r-inla.org` & `inlabru.org`)

$$
\begin{aligned}
\widetilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) &\propto \left. \frac{p(\boldsymbol{\theta}) p(\boldsymbol{u} \mid \boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\theta})}{p_G(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta})} \right|_{\boldsymbol{u}=\widetilde{\boldsymbol{\mu}}(\boldsymbol{\theta})} \\
\widetilde{p}(\boldsymbol{u}_i \mid \boldsymbol{y}) &\propto \int p_{GG}(\boldsymbol{u}_i \mid \boldsymbol{y}, \boldsymbol{\theta}) \widetilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) \, d\boldsymbol{\theta}
\end{aligned}
$$

The main practical limiting factors for the INLA method are the number of latent variables and the number model parameters.

Integrated Nested Laplace Approximation (INLA) basics

1. Estimate the posterior mode for $p(\boldsymbol{\theta}|\boldsymbol{y})$ by optimisation of the approximation

$$\widetilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \left.\frac{p(\boldsymbol{\theta})p(\boldsymbol{u} \mid \boldsymbol{\theta})p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\theta})}{p_G(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta})}\right|_{\boldsymbol{u}=\widetilde{\boldsymbol{\mu}}(\boldsymbol{\theta})}$$

   where $p_G(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta})$ is a Gaussian approximation matching the low order derivatives at the mode of the exact conditional log-posterior for $\boldsymbol{u}$. (In a fully Gaussian model this is exact.)
   This is a Laplace approximation of $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.
2. Construct a numerical integration grid/scheme $(\boldsymbol{\theta}_k, w_k)$ for $\boldsymbol{\theta}$, where w_k are integration weights; this step also estimates the normalisation constant.
3. Construct $p_{GG}(\boldsymbol{u}_i \mid \boldsymbol{y}, \boldsymbol{\theta}_k)$ as Laplace approximations of the marginal conditional posterior densities, integrating out $\boldsymbol{u}_{-i} = \{u_j, j \neq i\}$.
4. Combine to form marginal posterior densities:

$$\widetilde{p}(\boldsymbol{u}_i \mid \boldsymbol{y}) \propto \sum_k p_{GG}(\boldsymbol{u}_i \mid \boldsymbol{y}, \boldsymbol{\theta}_k)\, \widetilde{p}(\boldsymbol{\theta}_k \mid \boldsymbol{y})\, w_k$$

   This is a Gaussian mixture distribution.

Example: Point process data

Log-Gaussian Cox processes

Point intensity:

$$\lambda(\boldsymbol{s}) = \exp\left(\sum_i b_i(\boldsymbol{s})\beta_i + u(\boldsymbol{s})\right)$$

Inhomogeneous Poisson process log-likelihood:

$$\ln p(\{\boldsymbol{y}_k\} \mid \boldsymbol{\lambda}) = |D| - \int_D \lambda(\boldsymbol{s})\, d\boldsymbol{s} + \sum_{k=1}^{N} \ln \lambda(\boldsymbol{y}_k)$$

The likelihood can be approximated numerically, e.g.

$$\int_D \lambda(\boldsymbol{s})\, d\boldsymbol{s} \approx \sum_{j=1}^{n} \lambda(\boldsymbol{s}_j) w_j,$$

where s_j are mesh nodes, and $w_j = \langle \psi_j, 1 \rangle_D$

Example: Point process data (cont)

Discretised field and likelihood:

$$\lambda(\boldsymbol{s}) = \exp\left(\sum_i b_i(\boldsymbol{s})\beta_i + \sum_j \psi_j(\mathbf{s})u_j\right)$$

$$\ln p(\{\boldsymbol{y}_k\} \mid \boldsymbol{\lambda}) \approx |D| - \sum_{j=1}^{n} \lambda(\boldsymbol{s}_j)w_j + \sum_{k=1}^{N} \ln \lambda(\boldsymbol{y}_k)$$

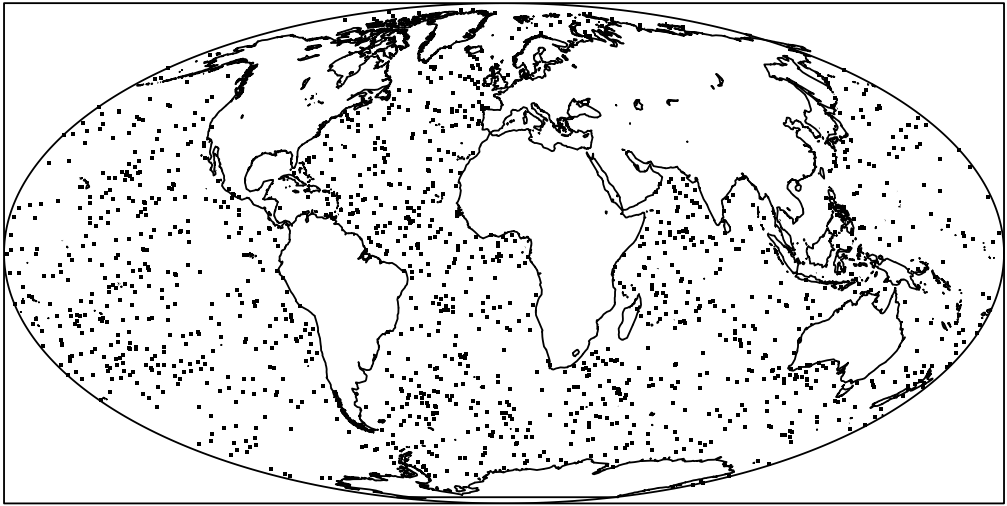
Then, with $\boldsymbol{\lambda_D} = \big[\lambda(s_i)\big]$, $\boldsymbol{A}_D = \big[\psi_j(s_i)\big]$, and $\boldsymbol{A}_y = \big[\psi_j(y_i)\big]$,

$$\nabla_{\boldsymbol{u}} \ln p(\{\boldsymbol{y}_k\} \mid \boldsymbol{\lambda}) \approx -\boldsymbol{A}_D^\top \operatorname{diag}(\boldsymbol{w})\boldsymbol{\lambda}_D + \boldsymbol{A}_y^\top \mathbf{1}$$

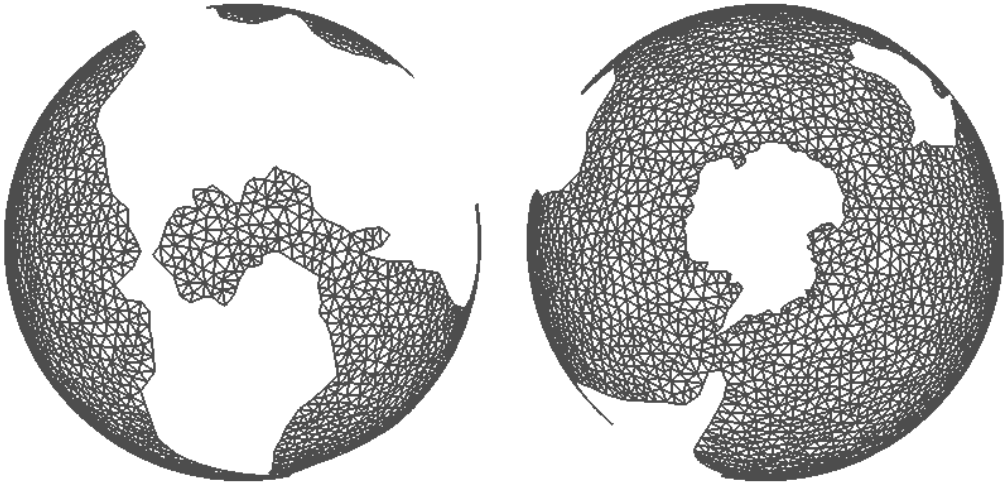
$$\nabla^2_{\boldsymbol{u}} \ln p(\{\boldsymbol{y}_k\} \mid \boldsymbol{\lambda}) \approx -\boldsymbol{A}_D^\top \operatorname{diag}(\boldsymbol{w}) \operatorname{diag}(\boldsymbol{\lambda}_D)\boldsymbol{A}_D$$

and similarly for ∇_β, ∇^2_β, and $\nabla_u \nabla_\beta$.

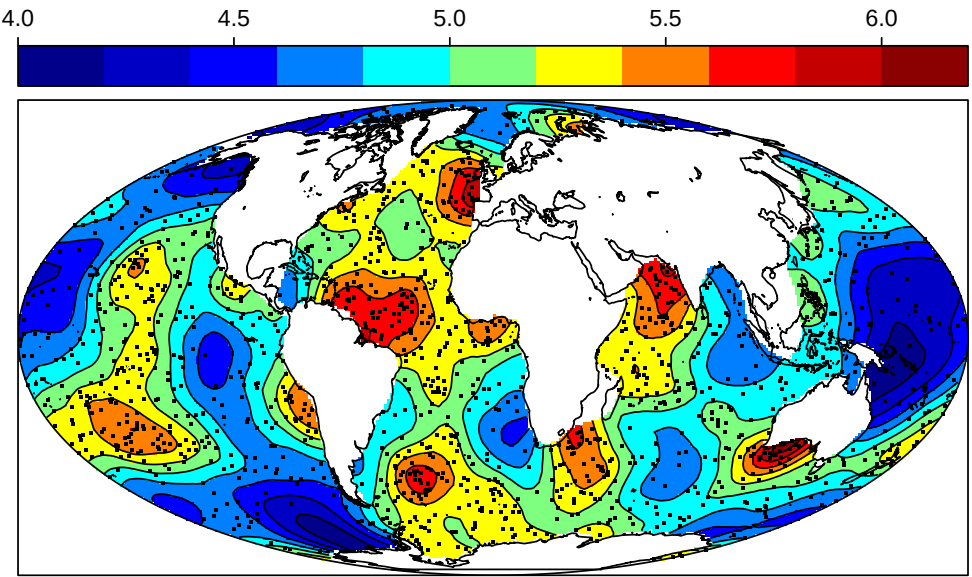
Concept illustration: rogue waves



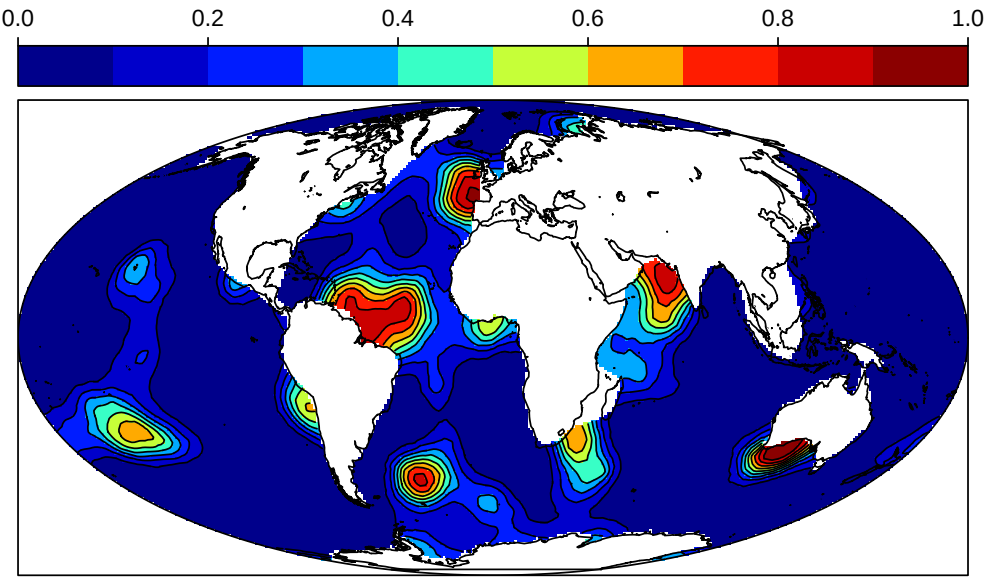
Mesh of the ocean subset of the globe



Posterior mean of the log-intensity



Marginal posterior probabilities for exceeding a threshold



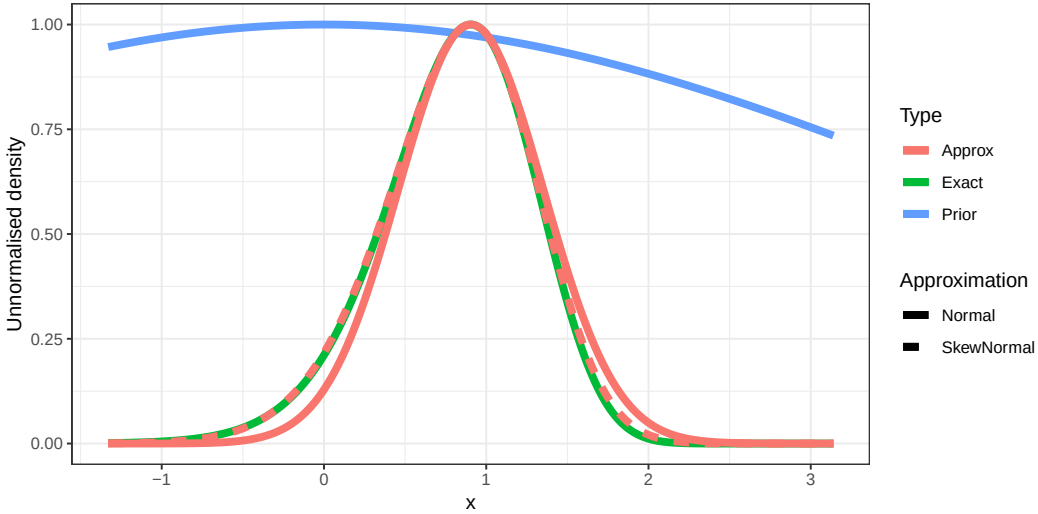
Bias and skewness improvement

- For skewed posteriors, the Normal approximation $p_{GG}(\cdot)$ at the mode is biased
- Can use higher order derivatives at the mode to find better approximations
- Example: Match 2nd and 3rd order derivatives of the log-posterior density to a skew-Normal distribution, at the posterior mode.
- The R-INLA implementation uses a different but related approach.

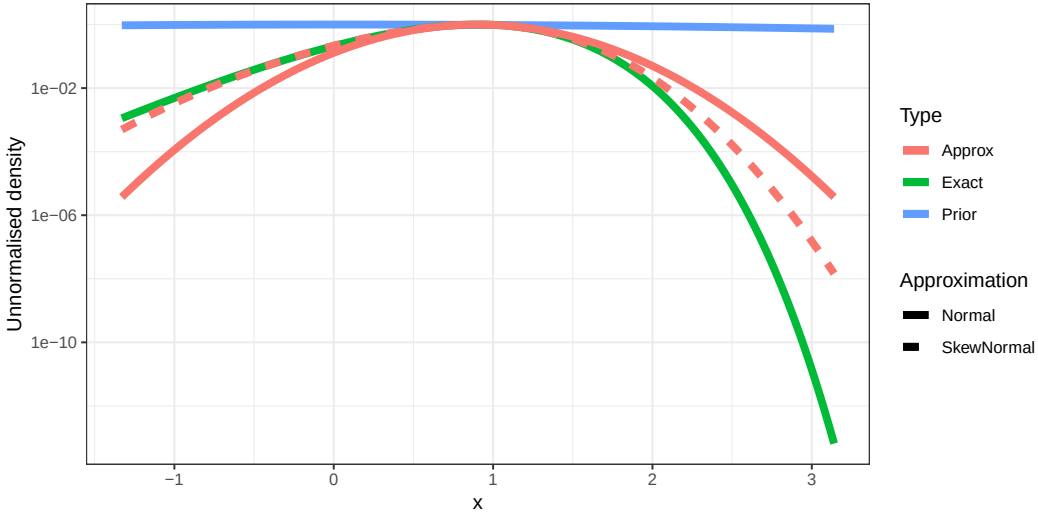
Skew-Normal distribution

Let $z = (x - m)/s$.
The skew-Normal density is defined by $p(x) = \frac{2}{s}\phi(z)\Phi(\alpha z)$, where $\alpha \in \mathbb{R}$ controls the skewness.
The first order derivative of the log-density is $-z + \frac{\alpha\phi(\alpha z)}{s\Phi(\alpha z)}$.
Higher order derivatives are straightforward (but tedious) to derive.

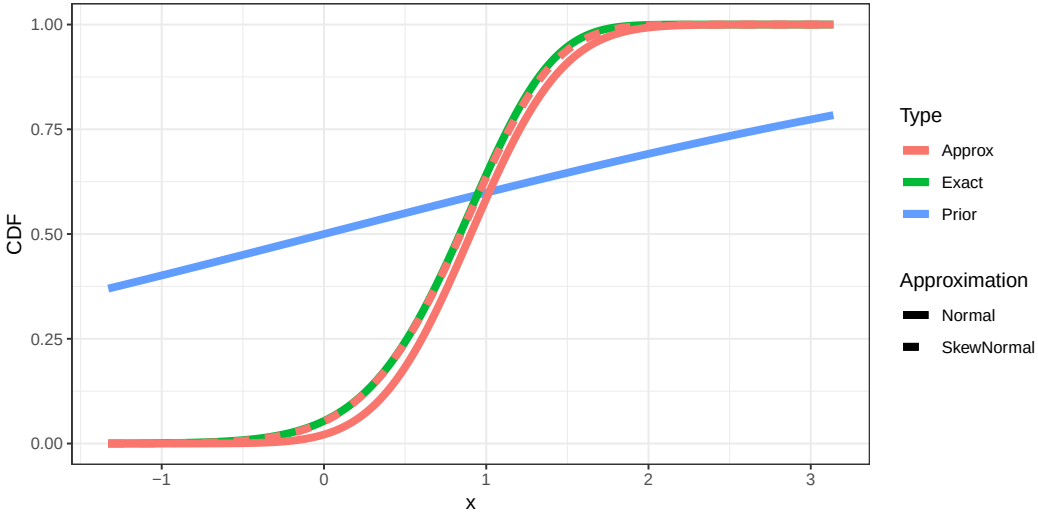
Densities


1.00
0.75
0.50
0.25
0.00
Unnormalised density
−1
0
1
2
3
x
Type
Approx
Exact
Prior
Approximation
Normal
SkewNormal

Log-densities



Cumulative distribution functions (CDF)



Bayesian method correctness assessment

The aim is to assess the correctness of the computed posterior distribution for some functional $h(\boldsymbol{\theta}, \boldsymbol{u})$.

For each $k = 1, \ldots, K$,

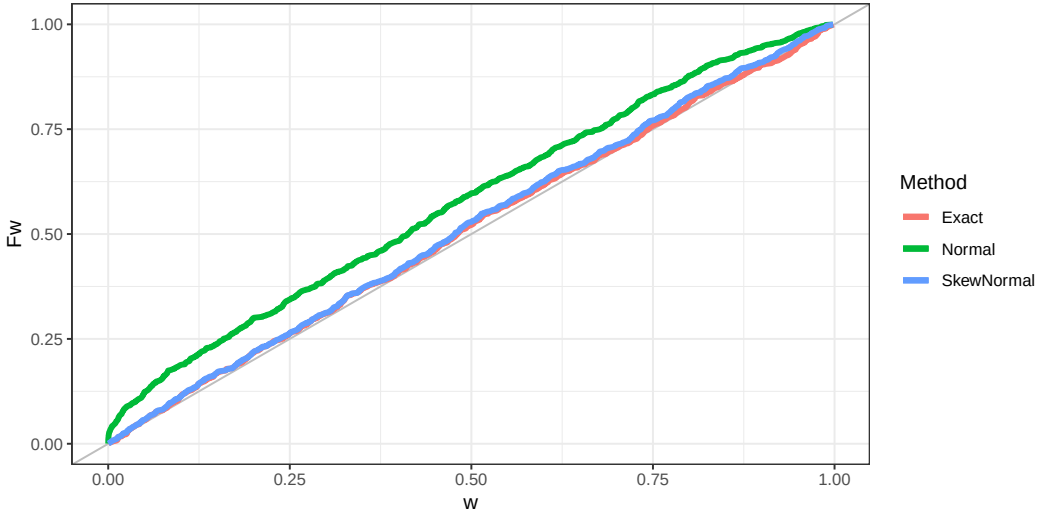
1. Sample
$$\boldsymbol{\theta}^{(k)} \sim p(\boldsymbol{\theta})$$
$$\boldsymbol{u}^{(k)} \sim p(\boldsymbol{u} \mid \boldsymbol{\theta}^{(k)})$$
$$\boldsymbol{y}^{(k)} \sim p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(k)}, \boldsymbol{u}^{(k)})$$
2. The method being assessed has posterior density approximation $\widehat{p}(h(\boldsymbol{\theta}, \boldsymbol{u})|\boldsymbol{y}^{(k)})$
3. Compute The CDF value $w^{(k)} = F_{\widehat{p}(h(\boldsymbol{\theta},\boldsymbol{u})|\boldsymbol{y}^{(k)})}(h(\boldsymbol{\theta}^{(k)}, \boldsymbol{u}^{(k)}))$

If the method recovers the correct posterior distributions, then $w^{(k)} \sim \mathsf{Unif}(0, 1)$, independent over $k = 1, \ldots, K$.

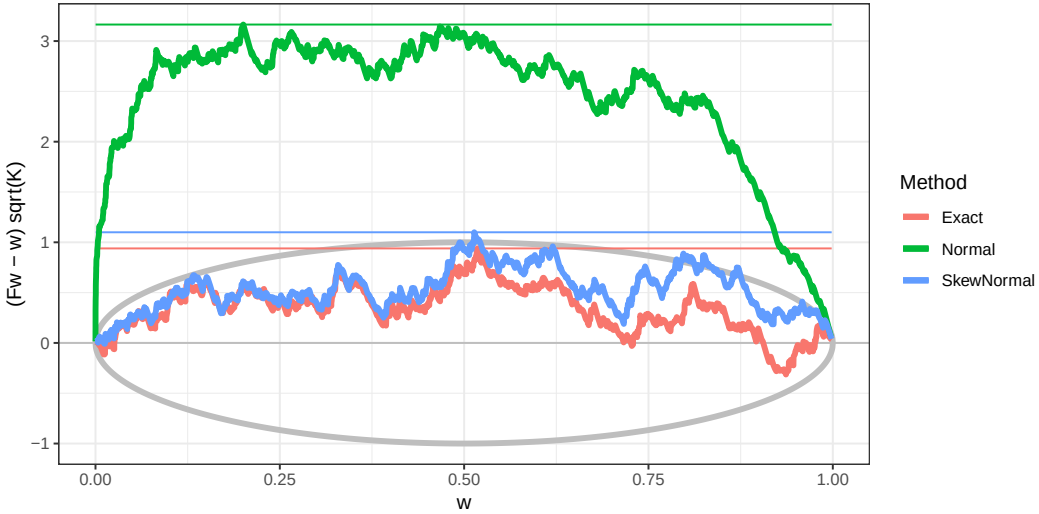
If the method does not recover the correct posterior distributions, then we expect to see some deviation from $\mathsf{Unif}(0, 1)$.

Example:
$$\begin{aligned} u &\sim \mathsf{N}(0, 4^2), \\ (y_i|u) &\sim \mathsf{Po}\left(\mathrm{e}^u\right), \quad \text{independent over } i = 1, \ldots, n = 5, \\ h(u) &= u. \end{aligned}$$

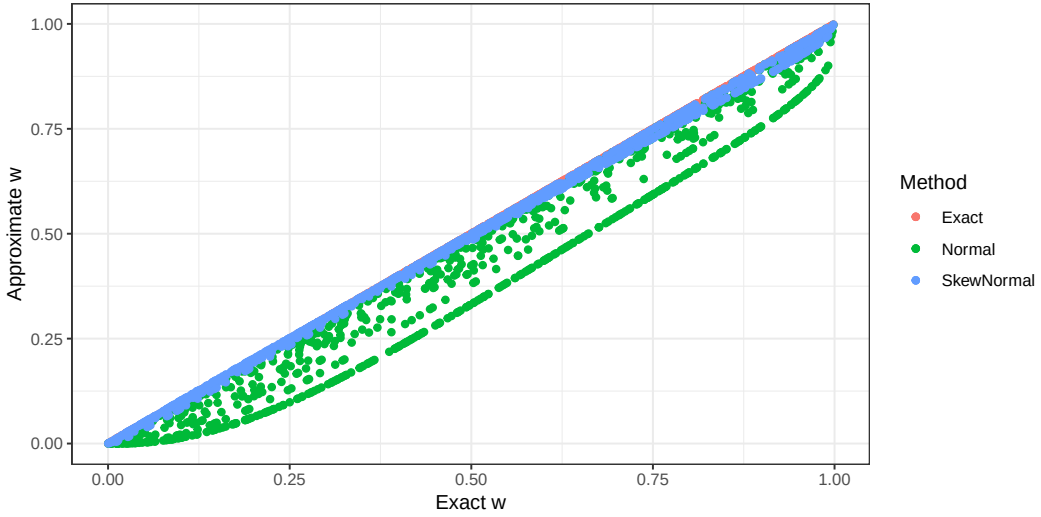
CDF comparison



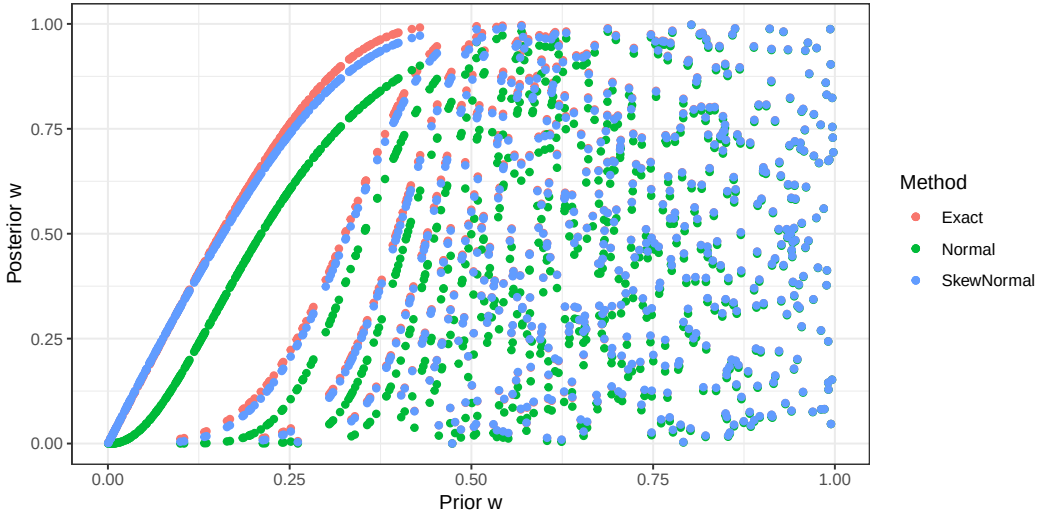
Kolmogorov-Smirnov diagnostic plots



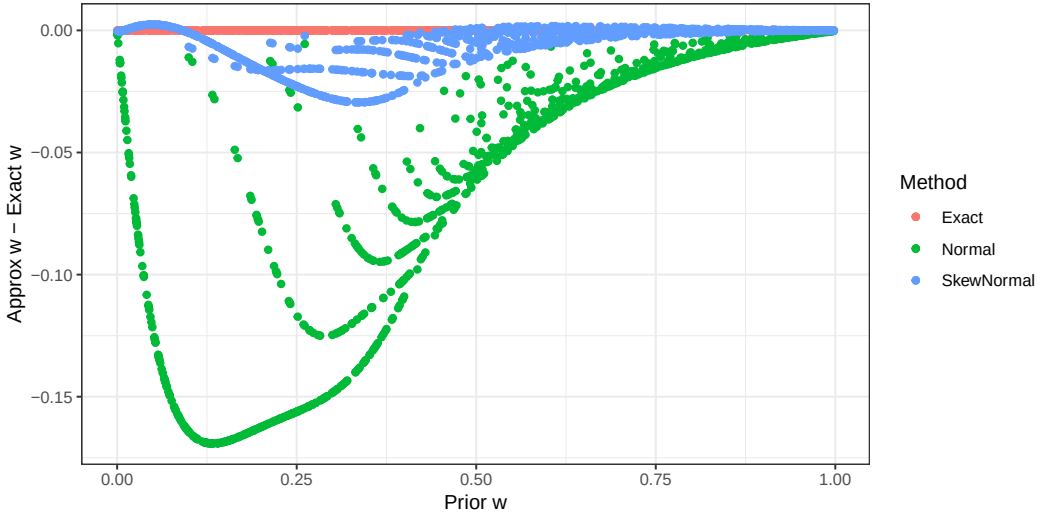
Exact vs Approximate w



Prior vs Posterior w



Prior vs Posterior w difference



Procedure for sampling based Bayesian methods

MCMC and other Monte Carlo methods do not provide CDF values, which then need to be estimated.

Posterior correctness assessment from samples (inspired by Talts et al, 2018)

Generate samples $(\boldsymbol{\theta}^{(k)}, \boldsymbol{u}^{(k)}, \boldsymbol{y}^{(k)})$ from the full model, as before.
For each $k = 1, \ldots, K$, generate J samples from $(\boldsymbol{\theta}^{(j|k)}, \boldsymbol{u}^{(j|k)}) \sim p(\boldsymbol{\theta}, \boldsymbol{u} \mid \boldsymbol{y}^{(k)})$.
Compute the approximate CDF value as an empirical CDF for the samples:

$$w^{(k)} = \frac{1}{J} \sum_{j=1}^{J} \mathbb{I}\left\{ h(\boldsymbol{\theta}^{(j|k)}, \boldsymbol{u}^{(j|k)}) \leq h(\boldsymbol{\theta}^{(k)}, \boldsymbol{u}^{(k)}) \right\} - \frac{1}{2J}$$

which is a normalised order statistic.

Notes:

- The assessment approach assumes we can sample *exactly* (and independently) from the prior model.
- MCMC methods capable of posterior samples are not necessarily good at generating from the prior.
- The null distribution for the K-S test now depends on both K and J, as well as the dependence between the Monte Carlo samples.

Model vs method vs implementation

- The Bayesian method correctness assessment method specifically targets the implementation of a method
- *Model assessment* based on output from a method implementation is meaningless if we we don't have some trust in the method and implementation
- Information criteria based on the full model likelihood are popular but are often hard to interpret
- Probabilistic predictions can be easier to interpret, and are often cheap to compute (in particular if one is already doing expensive Bayesian inference)
- We let F denote the CDF of a probabilistic prediction of an observation y
- The context can be cross-validation or estimation/validation/test data splits

Scores

- We want to quantify how well our predictions represent the test data.
- We define *scores* $S(F, y)$ that in some way measure how well the prediction F matched the actual value, y.
- The scores defined here are *negatively oriented*, meaning that the *lower the score, the better*.

Squared errors and log-likelihood scores

- Squared Error (SE): $S_{\text{SE}}(F, y) = (y - \widehat{y}_F)^2$,
  where $\widehat{y}_F$ is a point estimate under F, e.g. the expectation μ_F.
- Logarithmic/Ignorance score (LOG/IGN): $S_{\text{LOG}}(F, y) = -\log f(y)$,
  where $f(\cdot)$ is the predictive density.
- Dawid-Sebastiani (DS): $S_{\text{DS}}(F, y) = \frac{(y-\mu_F)^2}{\sigma_F^2} + \log(\sigma_F^2)$.

Score expectations and proper scoring rules

- What functions of the predictive distributions are useful scores?
- We want to reward accurate (unbiased) and precise (small variance) predictions, but not at the expense of understating true uncertainty.
- First, we define the expectation of a score under a true distribution G as

$$S(F, G) = \mathsf{E}_{y\sim G}[S(F, y)]$$

Proper scores/scoring rules

A negatively oriented score is *proper* if it fulfils

$$S(F, G) \geq S(G, G).$$

A proper score that has equality of the expectations *only* when F and G are the same, $F(\cdot) \equiv G(\cdot)$, is said to be *strictly proper*.

The practical interpretation of this is that a proper score does not reward cheating; stating a lower (or higher) forecast/prediction uncertainty will not, on average, give a better score than stating the truth.

Absolute error and CRPS

Absolute error and Continuous Ranked Probability Score

- Absolute Error (AE): $S_{\text{AE}}(F, y) = |y - \widehat{y}_F|$, where $\widehat{y}_F$ is a point estimate under F, e.g. the *median* $F^{-1}(1/2)$.
- CRPS: $S_{\text{CRPS}}(F, y) = \int_{-\infty}^{\infty} \left[\mathbb{I}(y \leq x) - F(x)\right]^2 \, \mathrm{d}x$

Average scores

Average score

Given a collection of prediction/truth pairs, $\{(F_i, y_i), i = 1, \ldots, n\}$, define the *average* or *mean* score:

$$\overline{S}(\{(F_i, y_i), i = 1, \ldots, n\}) = \frac{1}{n} \sum_{i=1}^{n} S(F_i, y_i)$$

- When comparing prediction quality, we often look at the difference in average scores across the test data set.
- For modern, complex models with explicit spatial and temporal model components, the *pairwise* differences may be useful: For two prediction methods, F and F',

$$S_i^{\Delta}(F_i, F_i', y_i) = S(F_i, y_i) - S(F_i', y_i)$$

  We can have $\overline{S}^{\Delta} \approx 0$ at the same time as all $|S_i^{\Delta}| \gg 0$, if the two models/methods are both good, but e.g. at different spatial locations.
- How can we assess whether the score differences are indistinguishable?

How good are confidence/prediction interval procedures?

Tradeoffs for CIs

Desired properties for methods generating CIs for a quantity Y:

1. Appropriate *coverage* under the true distribution, G: $\mathrm{P}_G(Y \in CI_F) \geq 1 - \alpha$
2. Narrow intervals

- A wide prediction F helps with 1 but makes 2 difficult
- A narrow prediction F helps with 2 but makes 1 difficult

A proper score for interval predictions

The *Interval Score* For a CI (L_F, U_F) is defined by

$$S_{\mathrm{INT}}(F, y) = U_F - L_F + \frac{2}{\alpha}(L_F - y)\mathbb{I}(y < L_F) + \frac{2}{\alpha}(y - U_F)\mathbb{I}(y > U_F)$$

It is a proper scoring rule, consistent for equal-tail error probability intervals:
$S(F, G)$ is minimised for the narrowest CI that has expected coverage $1 - \alpha$.

Proper scores

$$
\begin{aligned}
S_{\mathsf{SE}}(F, G) &= \mathsf{E}_{y\sim G}[S_{\mathsf{SE}}(F, y)] = \mathsf{E}_{y\sim G}[(y - \mu_F)^2] = \mathsf{E}_{y\sim G}[(y - \mu_G + \mu_G - \mu_F)^2] \\
&= \mathsf{E}_{y\sim G}[(y - \mu_G)^2 + 2(y - \mu_G)(\mu_G - \mu_F) + (\mu_G - \mu_F)^2] \\
&= \mathsf{E}_{y\sim G}[(y - \mu_G)^2] + 2(\mu_G - \mu_F)\mathsf{E}_{y\sim G}[y - \mu_G] + (\mu_G - \mu_F)^2 \\
&= \sigma_G^2 + (\mu_G - \mu_F)^2
\end{aligned}
$$

This is minimised when $\mu_F = \mu_G$. Therefore $S_{\mathsf{SE}}(F, G) \geq S_{\mathsf{SE}}(G, G) = \sigma_G^2$, so the score is proper. Is it strictly proper?

$$
\begin{aligned}
S_{\mathsf{DS}}(F, G) = \mathsf{E}_{y\sim G}[S_{\mathsf{DS}}(F, y)] &= \frac{\mathsf{E}_{y\sim G}[(y - \mu_F)^2]}{\sigma_F^2} + \log(\sigma_F^2) \\
&= \frac{\sigma_G^2 + (\mu_G - \mu_F)^2}{\sigma_F^2} + \log(\sigma_F^2)
\end{aligned}
$$

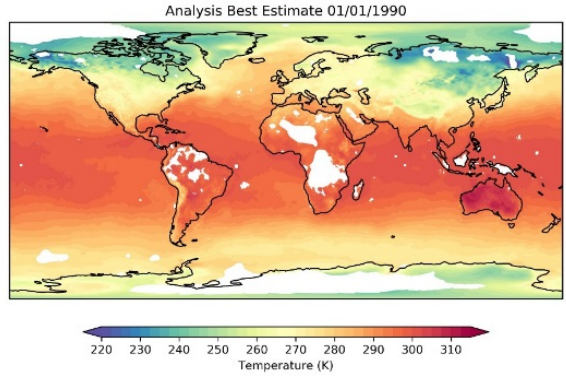
This is minimised when $\mu_F = \mu_G$ and $\sigma_F = \sigma_G$. Therefore $S_{\mathsf{DS}}(F, G) \geq S_{\mathsf{DS}}(G, G) = 1 + \log(\sigma_G^2)$, so the score is proper. Is it strictly proper?

Part 3: Lessons from the EUSTACE project

EUSTACE ANALYSIS

Combines in-situ and satellite data sources to derive daily air temperatures across the globe with quantified uncertainties.

- Daily mean air temperature (2 m) estimates from the mid-late 19th century at ¼ degree resolution.
- Observational dataset for use in climate monitoring, services and research.
  - Quantify bias and uncertainty arising from observational sampling (in space and time);
  - Quantify uncertainty from instrumental effects/network changes.
- Higher resolution daily gridded analyses for regional climate
  - Combine in situ and remote sensing data to support high resolution analysis.
  - Absolute temperature rather than anomaly product.



Met Office
Hadley Centre

EUSTACE

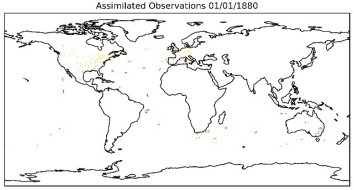
OBSERVATIONS

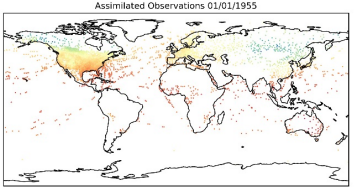
In situ air temperature:

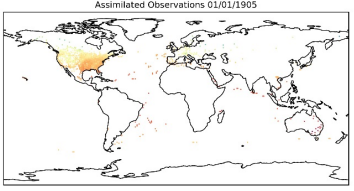
- EUSTACE station dataset (UBERN) (GHCN-D, ECA&D, ISTI, DECADE, ERA-CLIM)
- HadNMAT-2 ship air temperatures (NOCS/Met Office)

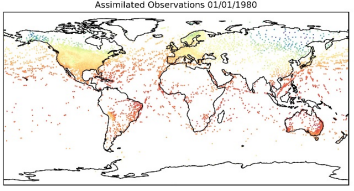
Satellite skin temperature derived air temperature:

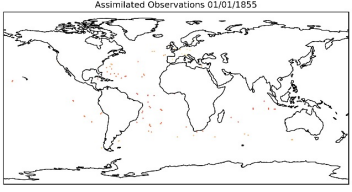
- Marine: ATSR (ESA CCI SST)
- Land: MODIS (USGS/NASA via ESA GlobTemperature)
- Ice: AVHRR (NOAA/FP7 NACLIM)

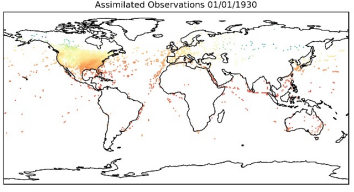


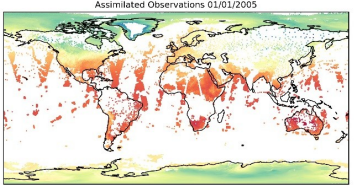












Met Office
Hadley Centre

Statistical model and method building blocks

Basic system components

- Temperature *processes on different spatial and temporal scales*
  - Seasonal
  - Slow climate processes
  - Medium-scale variability
  - Daily
- *Vast model size* ($\sim 10^{11}$ unknowns); need computationally efficient tools
- Hierarchical statistical model structure based on Gaussian processes
  - Stochastic PDEs translates to sparse precisions in *Gaussian Markov random fields*
- *Propagated uncertainty* via a Bayesian approach
  - Dependence structure parameters
  - Spatio-temporal process priors
  - Observation models; Multiple *observation sources*, with complex error *uncertainty structure*
- Goals:
  - a *best estimate*,
  - a *collection of samples*, and
  - more precise (and accurate) *uncertainty estimates*.

Matérn driven heat equation on the sphere

The iterated heat equation is a simple non-separable space-time SPDE family:

$$(\kappa^2 - \Delta)^{\gamma/2} \left[\phi \frac{\partial}{\partial t} + (\kappa^2 - \Delta)^{\alpha/2} \right]^{\beta} x(\mathbf{s}, t) = \mathcal{W}(\mathbf{s}, t)/\tau$$

For constant parameters, $x(\mathbf{s}, t)$ has spatial Matérn covariance (for each t).

Discrete domain Gaussian Markov random fields (GMRFs)

$\boldsymbol{x} = (x_1, \ldots, x_n) \sim \mathsf{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$ is Markov with respect to a neighbourhood structure $\{\mathcal{N}_i, i = 1, \ldots, n\}$ if $Q_{ij} = 0$ whenever $j \neq \mathcal{N}_i \cup i$.

- Project the SPDE solution space onto local basis functions:
  random Markov dependent basis weights (Lindgren et al, 2011).

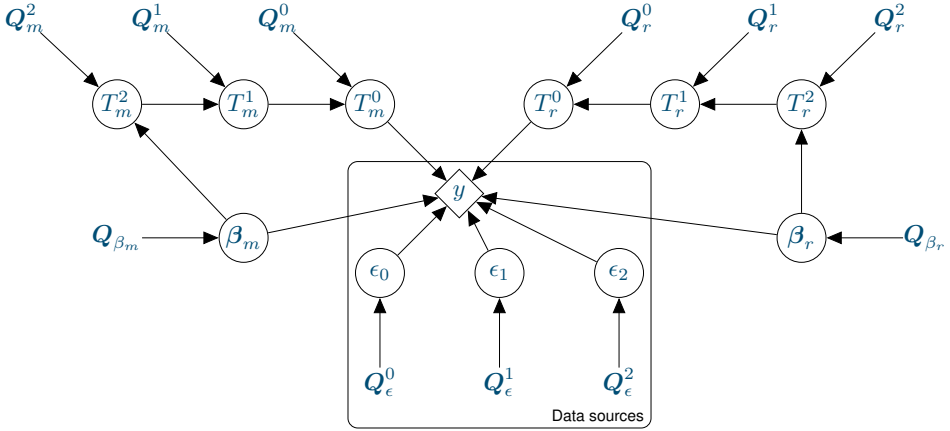
A finite element approximation has structure

$$x(\boldsymbol{s}, t) = \sum_{i,j} \psi_i^{[s]}(\boldsymbol{s}) \psi_j^{[t]}(t) x_{ij}, \quad \boldsymbol{x} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1}), \quad \boldsymbol{Q} = \sum_{k=0}^{\alpha+\beta+\gamma} \boldsymbol{M}_k^{[t]} \otimes \boldsymbol{M}_k^{[\mathbf{s}]}$$

even, e.g., if the spatial scale parameter κ is spatially varying.

Partial hierarchical representation

Observations of *mean*, *max*, *min*. Model *mean* and *range*.



Conditional specifications, e.g.

$$(T_m^0 | T_m^1, \boldsymbol{Q}_m^0) \sim \mathsf{N}\left(T_m^1, \boldsymbol{Q}_m^{0\,-1}\right)$$

$$T_r^0 = \exp(T_r^1)\, G^{-1}\left[U_r^0(\mathbf{s}, t)\right], \quad U_r^0 \sim \mathsf{N}\left(\mathbf{0}, \boldsymbol{Q}_r^{0\,-1}\right)$$

Standardised observation uncertainty models

- Each data source may have complicated dependence structure
- To facilitate information blending, use a common error term structure

Common satellite derived data error model framework

The observational&calibration errors are modelled as three error components:

- independent (ϵ_0),
- spatially and/or temporally correlated (ϵ_1), and
- systematic (ϵ_2),

with distributions determined by the uncertainty information from satellite calibration models.
E.g., $y_i = T_m(\mathbf{s}_i, t_i) + \epsilon_0(\mathbf{s}_i, t_i) + \epsilon_1(\mathbf{s}_i, t_i) + \epsilon_2(\mathbf{s}_i, t_i)$

In practice, each data source might have several different components of each type; independent components can be merged, but not necessarily correlated or systematic components.

Station observation&homogenisation model

Daily means

For station k at day t_i,

$$y_m^{k,i} = T_m(\mathbf{s}_k, t_i) + \sum_{j=1}^{J_k} H_j^k(t_i) e_m^{k,j} + \epsilon_m^{k,i},$$

where $H_j^k(t)$ are temporal step functions, $e_m^{k,j}$ are latent bias variables, and $\epsilon_m^{k,i}$ are independent measurement and discretisation errors.

Daily mean/max/min

For station k at day t_i,

$$y_m^{k,i} = T_m(\mathbf{s}_k, t_i) + \widetilde{H}_m^k(t_i) + \epsilon_m^{k,i},$$

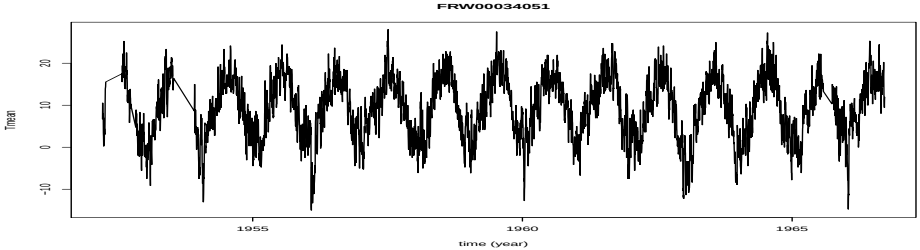
$$y_x^{k,i} = T_m(\mathbf{s}_k, t_i) + \frac{\exp[\widetilde{H}_r^k(t_i)]}{2} T_r(\mathbf{s}_k, t_i) + \epsilon_x^{k,i},$$

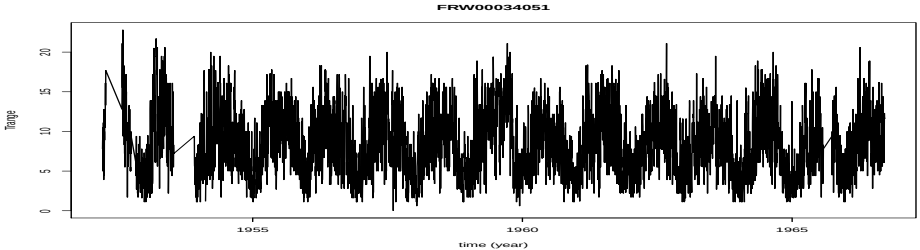
$$y_n^{k,i} = T_m(\mathbf{s}_k, t_i) - \frac{\exp[\widetilde{H}_r^k(t_i)]}{2} T_r(\mathbf{s}_k, t_i) + \epsilon_n^{k,i},$$

where $\widetilde{H}_\cdot^k$ are the total bias correction variables for each observation.

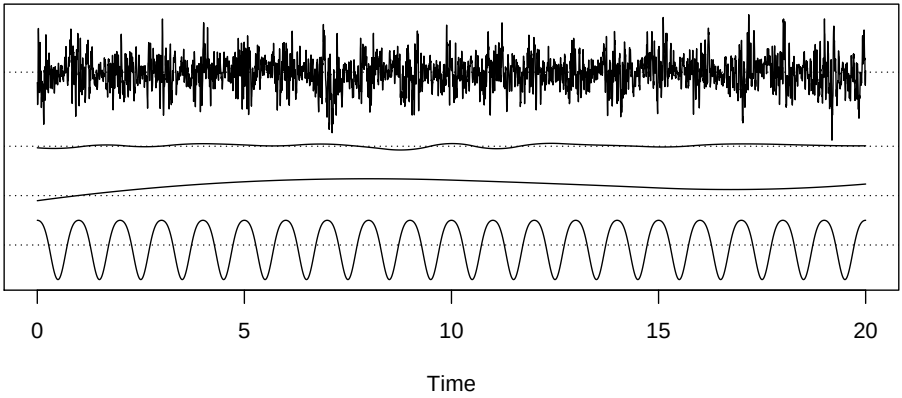
Observed data

Observed daily T_{mean} and T_{range} for station FRW00034051



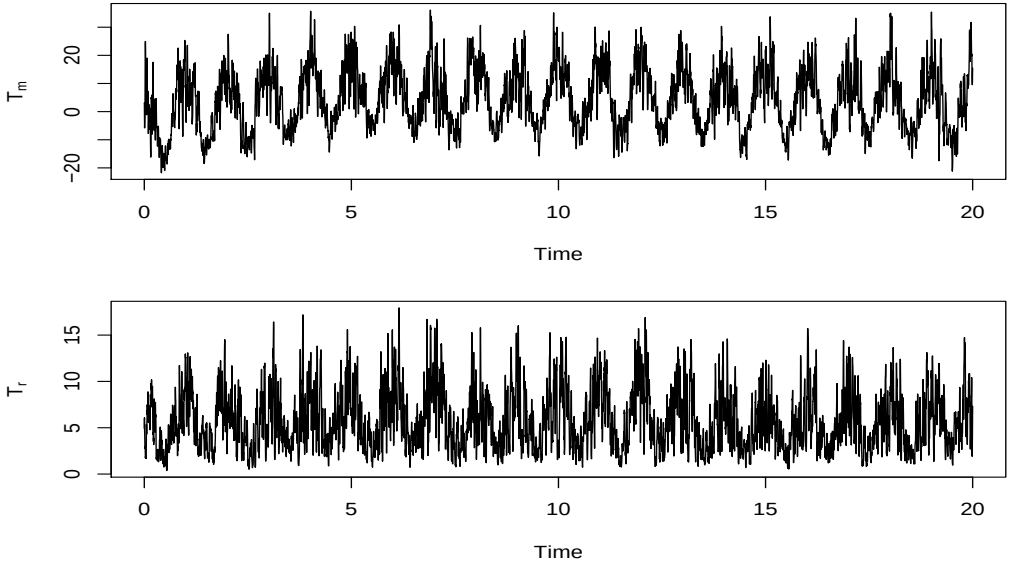


Multiscale model component samples



Combined model samples for T_m and T_r

(Proof of concept; no actual data was involved in this figure)



Linearised inference

All Spatio-temporal latent random processes combined into $\boldsymbol{x} = (\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{b})$, with joint expectation $\boldsymbol{\mu}_x$ and precision $\boldsymbol{Q}_x$:

$$(\boldsymbol{x} \mid \boldsymbol{\theta}) \sim \mathsf{N}(\boldsymbol{\mu}_x, \boldsymbol{Q}_x^{-1}) \quad \text{(Prior)}$$

$$(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) \sim \mathsf{N}(h(\boldsymbol{A}\boldsymbol{x}), \boldsymbol{Q}_{y|x}^{-1}) \quad \text{(Observations)}$$

$$p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}) \propto p(\boldsymbol{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) \quad \text{(Conditional posterior)}$$

Non-linear and/or non-Gaussian observations

For a non-linear $h(\boldsymbol{A}\boldsymbol{x})$ with Jacobian $\boldsymbol{J}$ at $\boldsymbol{x} = \widetilde{\boldsymbol{\mu}}$, iterate:

$$(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}) \overset{\text{approx}}{\sim} \mathsf{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{Q}}^{-1}) \quad \text{(Approximate conditional posterior)}$$

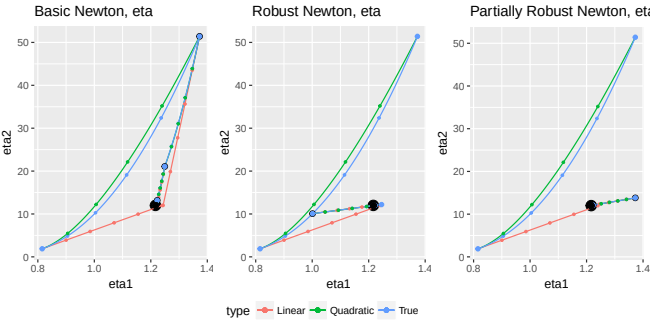
$$\widetilde{\boldsymbol{Q}} = \boldsymbol{Q}_x + \boldsymbol{J}^\top \boldsymbol{Q}_{y|x} \boldsymbol{J}$$

$$\widetilde{\boldsymbol{\mu}}' = \widetilde{\boldsymbol{\mu}} + a\widetilde{\boldsymbol{Q}}^{-1} \left\{ \boldsymbol{J}^\top \boldsymbol{Q}_{y|x} \left[\boldsymbol{y} - h(\boldsymbol{A}\widetilde{\boldsymbol{\mu}}) \right] - \boldsymbol{Q}_x (\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_x) \right\}$$

for some $a > 0$ chosen by line-search.

Iterative solutions for $\sim 10^{11}$ latent variables

- Nonlinear Newton iteration with robust line-search



- Preconditioned conjugate gradient (PCG) iteration for $\boldsymbol{Q}(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}) = \boldsymbol{r} = \boldsymbol{b} - \boldsymbol{Q}\widehat{\boldsymbol{\mu}}$
- Local and multiscale approximations for preconditioning: $\boldsymbol{M}^{-1}\boldsymbol{Q} \approx \boldsymbol{I}$
- Sampling with PCG: $\boldsymbol{Q}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}) = \boldsymbol{L}\boldsymbol{w}$
  Requires only a rectangular pseudo-Cholesky factorisation $\boldsymbol{L}\boldsymbol{L}^{\top} = \boldsymbol{Q}$.
  Possible due to the kronecker product sum precision structure.

MULTI-SCALE ANALYSIS MODEL

Statistical model for temperature variations and different scales (space and time):

- **Climatological variation**: local seasonal cycle with effects of latitude, altitude and coastal influence.
- **Large-scale variation**: Slowly varying climatological mean temperature field. Station homogenisation.
- **Daily Local**: daily variability associated with weather. Satellite retrieval biases.

Simultaneously estimates observational biases of known bias structures:

- e.g. satellite biases, station homogenisation.

Processed on STFC's LOTUS cluster www.jasmin.ac.uk:

- Largest solves processed on 20 core/256GB RAM node.
- Highly parallel observation pre-processing.

Element	Resolution	N Variables
Seasonal	Bimonthly x 1° SPDE	245,772
Slow-scale*	5 year x 5° SPDE	107,604
Latitude	0.5° latitude SPDE	721
Altitude	(0.25° grid)	1
Coastal	(0.25° grid)	1
Grand mean	Analysis mean	1

Element	Resolution	N Variables
Large-scale	3 monthly x 5° SPDE	1,752,408
Station bias	NA	82,072

Element	Resolution	N Variables per day
Daily local	~0.5 degree SPDE	162,842
Satellite bias (marine)	Global	1
Satellite bias (land)	Global + 2.5 degree SPDE	1 + 40,962
Satellite bias (ice)	Hemispheric + 2.5 degree SPDE*	2 + 40,962

Met Office Hadley Centre

EUSTACE

ITERATIVE SOLUTION

Condition on expected value of other components

(1) Climatology:
Global mean T
Seasonal cycle
Latitude
Altitude
Water fraction

$\boldsymbol{T}^{clim}(\boldsymbol{s},t)$

(2) Large-scale:
Multi-annual SPDE
Multi-month SPDE
Station biases

$\boldsymbol{T}^{large}(\boldsymbol{s},t)$

(3) Daily local:
Daily fields
Satellite land bias
Satellite marine bias
Satellite ice bias

$\boldsymbol{T}^{local}(\boldsymbol{s},t)$

Model Factorisation

The model is factorised into $m = 1, \ldots, 3$ components that are estimated interatively, substituting $\tilde{\boldsymbol{y}}_m$ for $\boldsymbol{y}$:

$$\tilde{\boldsymbol{y}}_m = \boldsymbol{y} - \sum_{n \neq m} \boldsymbol{J}_n \boldsymbol{\mu}_{\boldsymbol{x}_n | \tilde{\boldsymbol{y}}_n}$$

Met Office Hadley Centre

EUSTACE

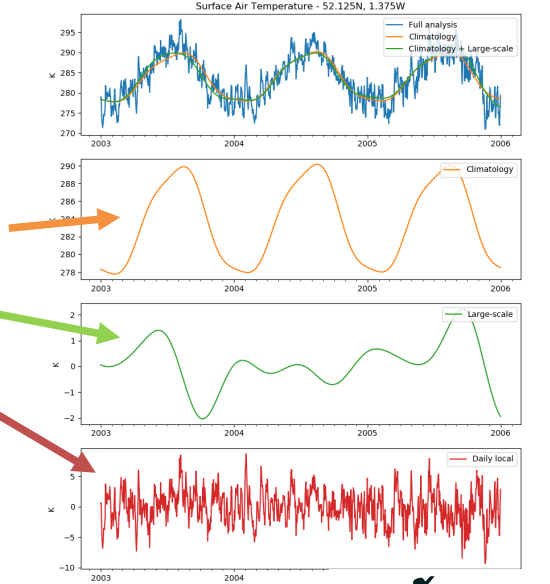
MULTI-SCALE ANALYSIS MODEL

Statistical model for temperature variations and different scales (space and time):

- **Climatological variation**: local seasonal cycle with effects of latitude, altitude and coastal influence.
- **Large-scale variation**: Slowly varying climatological mean temperature field.
- **Daily Local**: daily variability associated with weather.

Simultaneously estimates observational biases of known bias structures:

- e.g. satellite biases, station homogenisation.

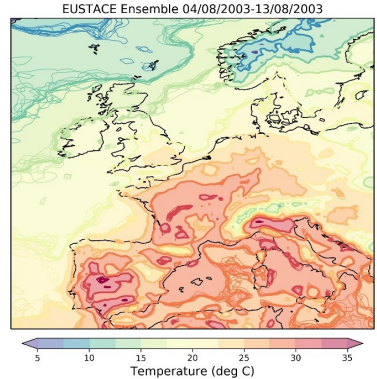


Met Office Hadley Centre

EUSTACE

ENSEMBLE ANALYSIS

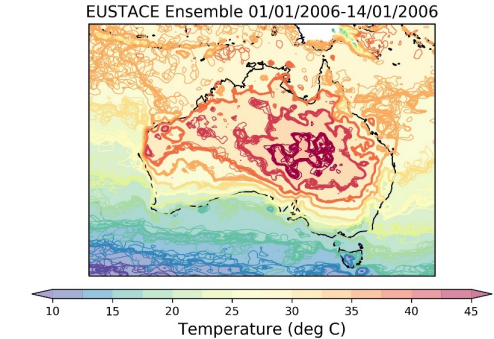
- Samples drawn from joint posterior distribution of temperature and bias variables.
- Temperature model samples projected onto analysis grid.
- Spatial/temporal correlation in analysis errors is encoded into the ensemble.
- Summary statistics can be derived from the ensemble. Expected value, total uncertainty and observation constraint information also available.



Met Office
Hadley Centre

ENSEMBLE ANALYSIS

- Samples drawn from joint posterior distribution of temperature and bias variables.
- Temperature model samples projected onto analysis grid.
- Spatial/temporal correlation in analysis errors is encoded into the ensemble.
- Summary statistics can be derived from the ensemble. Expected value, total uncertainty and observation constraint information also available.



Met Office
Hadley Centre

Variance calculations

Sparse partial inverse: Takahashi recursions postprocesses Cholesky

Takahashi recursions compute $\boldsymbol{S}$ such that $\boldsymbol{S}_{ij} = (\boldsymbol{Q}^{-1})_{ij}$ for all $Q_{ij} \neq 0$. Postprocessing of the (sparse) Cholesky factor. See `INLA::inla.qinv()`. Allows computing e.g. $\frac{\partial}{\partial \boldsymbol{\theta}} \log[\det(\boldsymbol{Q})] = \operatorname{tr}\left(\boldsymbol{Q}^{-1} \frac{\partial \boldsymbol{Q}}{\partial \boldsymbol{\theta}}\right)$.

Basic Rao-Blackwellisation of sample estimators

Let $\boldsymbol{x}^{(j)}$ be Gaussian samples and let $\boldsymbol{a}^\top \boldsymbol{x}$ be a linear combination of interest. For any subdomain $\Omega_k \subset \Omega$,

$$\mathsf{E}(\boldsymbol{a}^\top \boldsymbol{x}) = \mathsf{E}\left[\mathsf{E}(\boldsymbol{a}^\top \boldsymbol{x} \mid \boldsymbol{x}_{\Omega_k^*})\right] \approx \frac{1}{J} \sum_{j=1}^{J} \mathsf{E}(\boldsymbol{a}^\top \boldsymbol{x} \mid \boldsymbol{x}_{\Omega_k^*}^{(j)})$$

$$\begin{aligned}\mathsf{Var}(\boldsymbol{a}^\top \boldsymbol{x}) &= \mathsf{E}\left[\mathsf{Var}(\boldsymbol{a}^\top \boldsymbol{x} \mid \boldsymbol{x}_{\Omega_k^*})\right] + \mathsf{Var}\left[\mathsf{E}(\boldsymbol{a}^\top \boldsymbol{x} \mid \boldsymbol{x}_{\Omega_k^*})\right] \\ &\approx \mathsf{Var}(\boldsymbol{a}^\top \boldsymbol{x} \mid \boldsymbol{x}_{\Omega_k^*}^{(j)}) + \frac{1}{J} \sum_{j=1}^{J} \left[\mathsf{E}(\boldsymbol{a}^\top \boldsymbol{x} \mid \boldsymbol{x}_{\Omega_k^*}^{(j)}) - \mathsf{E}(\boldsymbol{a}^\top \boldsymbol{x})\right]^2\end{aligned}$$

Efficient if $\boldsymbol{a}\boldsymbol{a}^\top$ sparsity matches $\boldsymbol{S}_k$ on each subdomain:

$$\mathsf{Var}(\boldsymbol{a}^\top \boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{Q}^{-1} \boldsymbol{a} = \operatorname{tr}(\boldsymbol{Q}^{-1} \boldsymbol{a}\boldsymbol{a}^\top) = \sum_{ij} \left[\boldsymbol{Q}^{-1} \odot \boldsymbol{a}\boldsymbol{a}^\top\right]_{ij} = \sum_{ij} \left[\boldsymbol{S} \odot \boldsymbol{a}\boldsymbol{a}^\top\right]_{ij}$$

Preconditioning for e.g. conjugate gradient solutions

Solving $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{b}$ is equivalent to solving $\boldsymbol{M}^{-1}\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{M}^{-1}\boldsymbol{b}$. Choosing $\boldsymbol{M}^{-1}$ as an approximate inverse to $\boldsymbol{Q}$ gives a less ill-conditioned system. Only the *action* of $\boldsymbol{M}^{-1}$ is needed, e.g. one or more fixed point iterations:

Block Jacobi and Gauss-Seidel preconditioning

$$\text{Matrix split: } \boldsymbol{Q}_{x|y} = \boldsymbol{L} + \boldsymbol{D} + \boldsymbol{L}^\top$$

$$\text{Jacobi: } \boldsymbol{x}^{(k+1)} = \boldsymbol{D}^{-1}\left(-(\boldsymbol{L} + \boldsymbol{L}^\top)\boldsymbol{x}^{(k)} + \boldsymbol{b}\right)$$

$$\text{Gauss-Seidel: } \boldsymbol{x}^{(k+1)} = (\boldsymbol{L} + \boldsymbol{D})^{-1}\left(-\boldsymbol{L}^\top\boldsymbol{x}^{(k)} + \boldsymbol{b}\right)$$

Remark: Block Gibbs sampling for a GMRF posterior

With $\boldsymbol{Q} = \boldsymbol{Q}_{x|y}$, $\boldsymbol{b} = \boldsymbol{A}^\top\boldsymbol{Q}_\epsilon(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}_x)$ and $\widetilde{\boldsymbol{x}} = \boldsymbol{x} - \boldsymbol{\mu}_x$,

$$\widetilde{\boldsymbol{x}}^{(k+1)} = (\boldsymbol{L} + \boldsymbol{D})^{-1}\left(-\boldsymbol{L}^\top\widetilde{\boldsymbol{x}}^{(k)} + \boldsymbol{b} + \widetilde{\boldsymbol{L}}_{\boldsymbol{D}}\boldsymbol{w}\right), \quad \boldsymbol{w} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I})$$

Gauss-Seidel and Gibbs are both very inefficient on their own.

Overlapping block preconditioning

For ease of notation, write the two-level model $\boldsymbol{x}_0 = \boldsymbol{B}\boldsymbol{x}_1 + \text{fine scale variability}$ posterior precision as

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Q}_0 + \boldsymbol{A}^\top \boldsymbol{Q}_\epsilon \boldsymbol{A} & -\boldsymbol{Q}_0 \boldsymbol{B} \\ -\boldsymbol{B}^\top \boldsymbol{Q}_0 & \boldsymbol{Q}_1 + \boldsymbol{B}^\top \boldsymbol{Q}_0 \boldsymbol{B} \end{bmatrix}$$

Overlapping block preconditioning

Let $\boldsymbol{D}_k^\top$ be a restriction matrix to subdomain Ω_k, and let $\boldsymbol{W}_k$ be a diagonal weight matrix. Then a useful additive Schwartz preconditioner is

$$\boldsymbol{M}^{-1}\boldsymbol{x} = \sum_{k=1}^{K} \boldsymbol{W}_k \boldsymbol{D}_k (\boldsymbol{D}_k^\top \boldsymbol{Q} \boldsymbol{D}_k)^{-1} \boldsymbol{D}_k^\top \boldsymbol{W}_k \boldsymbol{x}$$

The domain overlap may need to be substantial:

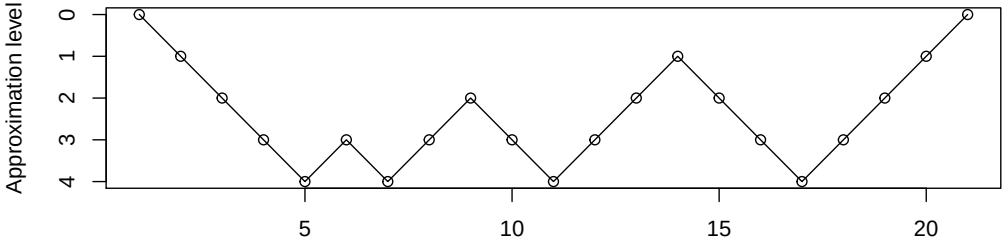
- Typical off-the-shelf preconditioning is aimed at at most 2nd order operators (Laplacian)
- In the example model, the spatial precision operator order is 6
- In a hierarchical triangle subdivision mesh, neighbouring hexagonal macro-domains overlap by 2 macro-triangles

Multigrid

Multigrid

Let $\boldsymbol{B}_c^\top$ be a projection matrix to a coarse approximative model. Then a basic multigrid step for $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{b}$ is

1. Apply high frequency preconditioner to get $\widehat{\boldsymbol{x}}_0$, let $\boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{Q}\widehat{\boldsymbol{x}}_0$
2. Project the problem to the coarser model: $\boldsymbol{Q}_c = \boldsymbol{B}_c^\top \boldsymbol{Q} \boldsymbol{B}_c, \boldsymbol{r}_c = \boldsymbol{B}_c^\top \boldsymbol{r}_0$
3. Apply multigrid to $\boldsymbol{Q}_c \boldsymbol{x}_c = \boldsymbol{r}_c$
4. Update the solution: $\widehat{\boldsymbol{x}}_1 = \widehat{\boldsymbol{x}}_0 + \boldsymbol{B}_c \widehat{\boldsymbol{x}}_c$
5. Apply high frequency preconditioner to get $\widehat{\boldsymbol{x}}_2$



The hierarchy of scales and preconditioning ($\boldsymbol{x}_0 = \boldsymbol{B}\boldsymbol{x}_1 + \text{fine scale variability}$):

Multiscale Schur complement approximation

Solving $\boldsymbol{Q}_{\boldsymbol{x}|\boldsymbol{y}}\boldsymbol{x} = \boldsymbol{b}$ can be formulated using two solves with the upper (fine) block $\boldsymbol{Q}_0 + \boldsymbol{A}^\top \boldsymbol{Q}_\epsilon \boldsymbol{A}$, and one solve with the *Schur complement*

$$\boldsymbol{Q}_1 + \boldsymbol{B}^\top \boldsymbol{Q}_0 \boldsymbol{B} - \boldsymbol{B}^\top \boldsymbol{Q}_0 \left(\boldsymbol{Q}_0 + \boldsymbol{A}^\top \boldsymbol{Q}_\epsilon \boldsymbol{A}\right)^{-1} \boldsymbol{Q}_0$$

By mapping the fine scale model onto the coarse basis used for the coarse model, we get an *approximate* (and sparse) Schur solve via

$$\begin{bmatrix} \widetilde{\boldsymbol{Q}}_B + \boldsymbol{B}^\top \boldsymbol{A}^\top \boldsymbol{Q}_\epsilon \boldsymbol{A}\boldsymbol{B} & -\widetilde{\boldsymbol{Q}}_B \\ -\widetilde{\boldsymbol{Q}}_B & \boldsymbol{Q}_1 + \widetilde{\boldsymbol{Q}}_B \end{bmatrix} \begin{bmatrix} \text{ignored} \\ \boldsymbol{x}_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \widetilde{\boldsymbol{b}} \end{bmatrix}$$

where $\widetilde{\boldsymbol{Q}}_B = \boldsymbol{B}^\top \boldsymbol{Q}_0 \boldsymbol{B}$.
The block matrix can be interpreted as the precision of a bivariate field on a common, coarse spatio-temporal scale, and the same technique applied to this system, with $\boldsymbol{x}_{1,1} = \boldsymbol{B}_{1|2}\boldsymbol{x}_{1,2} + \text{finer scale variability}$.

For realistic problems we need to combine all three techniques.

References (1/4)

- F. Lindgren, H. Rue and J. Lindström (2011),
  An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion), JRSSB, 73(4):423–498. Code available in `R-INLA`, see `http://r-inla.org/`
- Preliminary draft book chapter on GMRF computation basics (2015):
  `http://www.maths.ed.ac.uk/~flindgre/tmp/gmrf.pdf`
- R. Ingebrigtsen, F. Lindgren, I. Steinsland (2014), *Spatial models with explanatory variables in the dependence structure*, Spatial Statistics, 8:20–38. `http://dx.doi.org/10.1016/j.spasta.2013.06.002`
- G-A. Fuglstad, F. Lindgren, D. Simpson, H. Rue (2015),
  Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy, Statistica Sinica, 25:115–133. `http://arxiv.org/abs/1304.6949`
- Fabian E. Bachl, Finn Lindgren, David L. Borchers, and Janine B. Illian (2019)
  `inlabru`: an R package for Bayesian spatial modelling from ecological survey data, Methods in Ecology and Evolution, 10(6):760–766. `https://doi.org/10.1111/2041-210X.13168`
  CRAN package: `inlabru`, webpage `https://inlabru.org/`

References (2/4)

- Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, Andrew Gelman (Submitted on 18 Apr 2018)
  Validating Bayesian Inference Algorithms with Simulation-Based Calibration,
  http://arxiv.org/abs/1804.06788
- D. Bolin, F. Lindgren (2015),
  Excursion and contour uncertainty regions for latent Gaussian models,
  JRSSB, 77(1):85–106. http://arxiv.org/abs/1211.3946
  CRAN package: `excursions`
- H. Cramér and R. M. Leadbetter (2004, reprint of 1967 edition),
  Stationary and Related Stochastic Processes: Sample Function Properties and Their Applications,
  Dover Books on Mathematics.
- G. Lindgren (2012),
  Stationary Stochastic Processes: Theory and applications,
  Chapman & Hall/CRC, Texts in Statistical Science.

References (3/4)

- J. Besag (1974),
  Spatial interaction and the statistical analysis of latice systems (with discussion), JRSSB, 36(2):192–225.
- J. Besag and C. Kooperberg (1995),
  On conditional and intrinsic autoregressions, Biometrika, 82(4):733–746.
- J. Besag (1981),
  On a system of two-dimensional recurrence equations, JRSSB, 43(3):302–309.
- J. Besag and D. Mondal (2005),
  First-order intrinsic autoregressions and the de Wij process, Biometrika, 92(4):909–920.
- Y. A. Rozanov (1977),
  Markov random fields and stochastic partial differential equations, Mathematics of the USSR – Sbornik, 32(4):515–534
- Y. A. Rozanov (1977),
  On the paper "Markov random fields and stochastic partial differential equations", Mathematics of the USSR – Sbornik, 35(1):157–164
- Y. A. Rozanov (1982, Russian original 1980),
  Markov Random Fields, Springer Verlag, New York.

References (4/4)

- Y. A. Rozanov (2010, reprint of 1st ed. 1998),
  Random fields and stochastic partial differential equations, Springer Verlag, Mathematics and its Applications.
- R. J. Adler and J. Taylor (2007),
  Random Fields and Geometry, Springer Monographs in Mathematics
- P. Whittle (1954),
  On stationary processes in the plane, Biometrika, 41(3/4):434–449.
- P. Whittle (1963),
  Stochastic processes in several dimensions, Bull. Inst. Internat. Statist., 40:974–994.
- B. Matérn (1960),
  Spatial variation. Stochastic models and their application to some problems in forest surveys and other sampling investigations. (Swedish: Stokastiska modeller och deras tillämpning på några problem i skogstaxering och andra samplingundersökningar.), Stockholm University, PhD thesis in Mathematical Statistics. You may be lucky and find someone who has a copy!