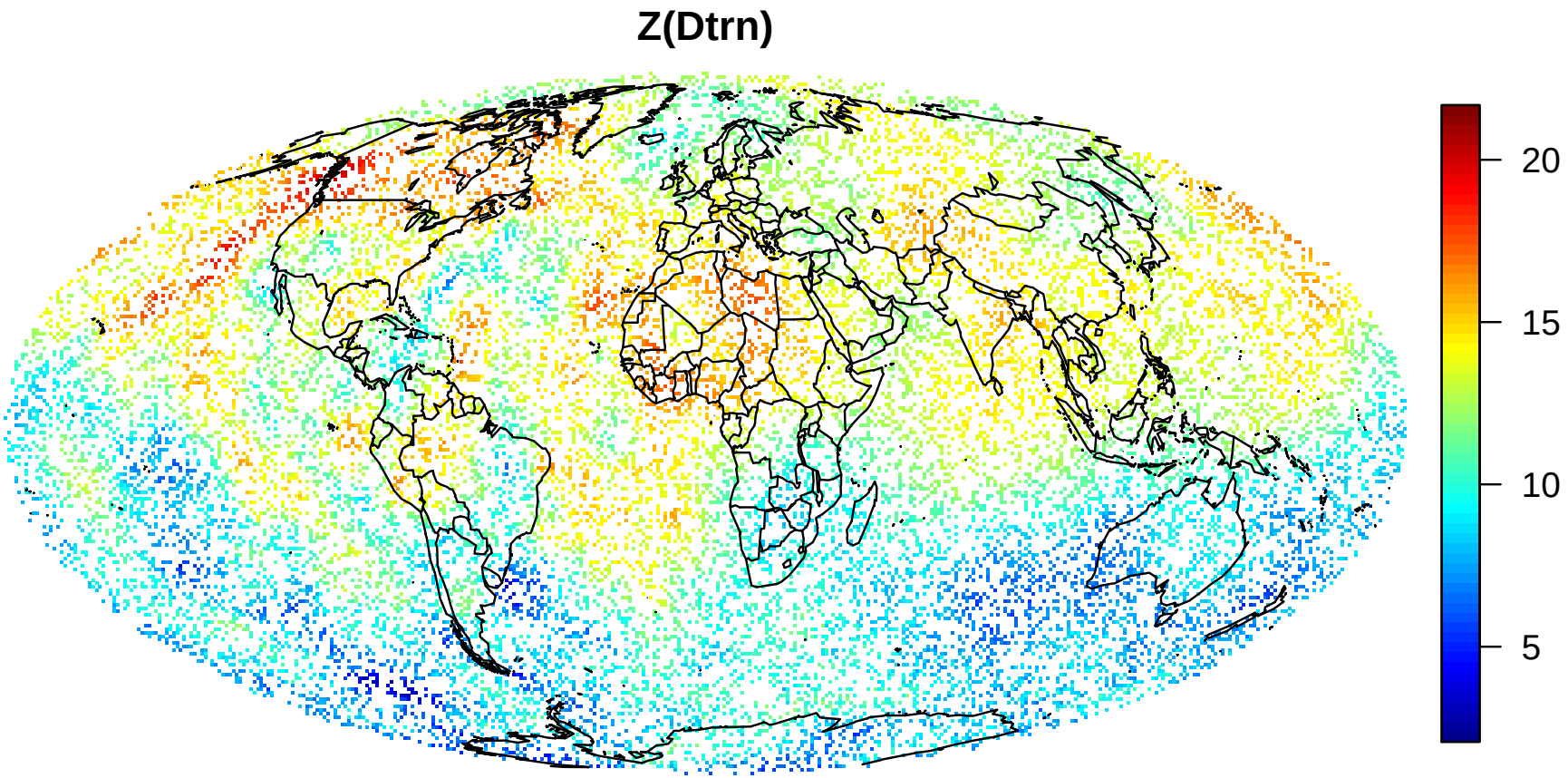
Big models for structured environmental big data

Finn Lindgren

London 2014-09-17

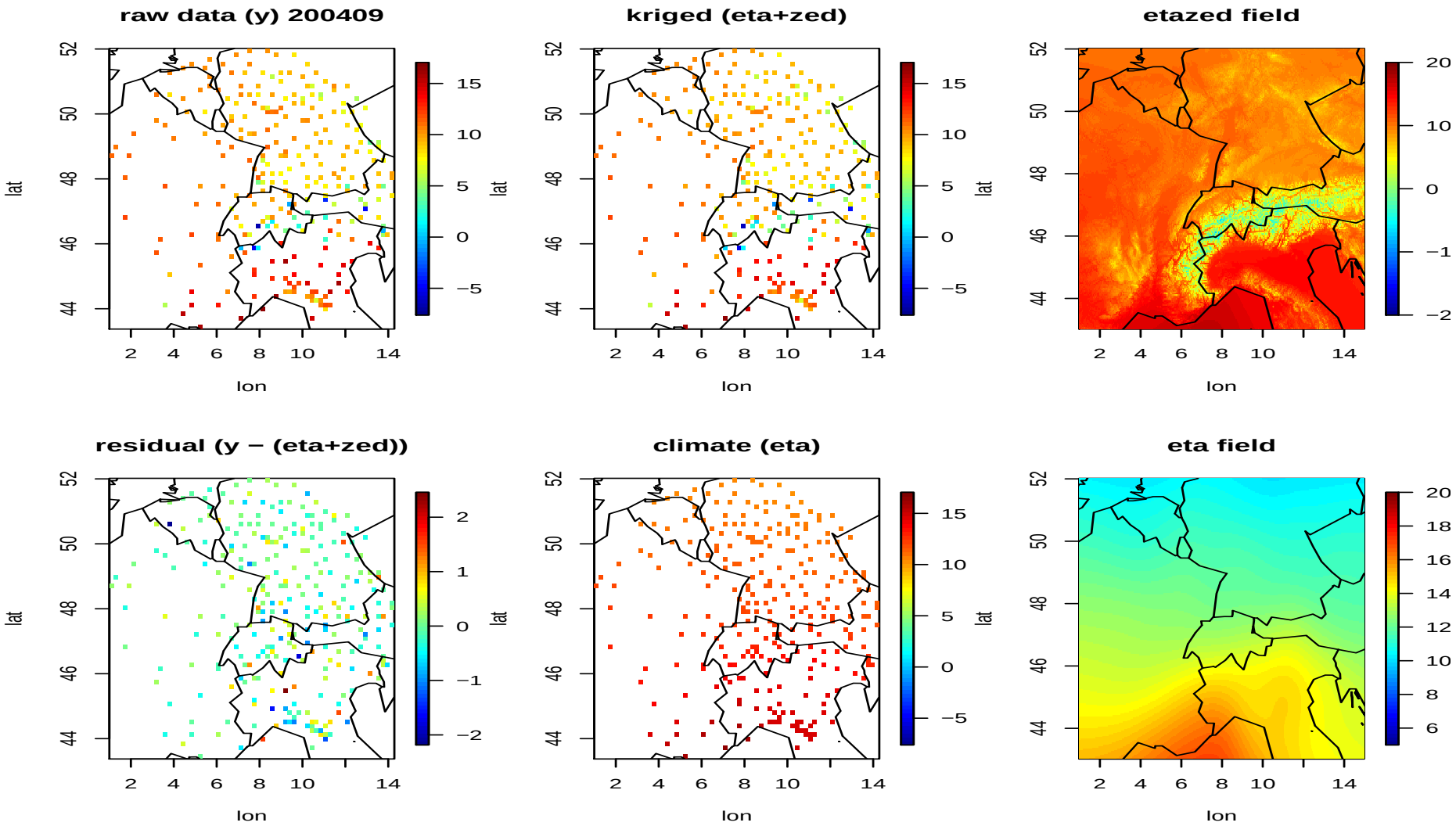
“Big” data



Synthetic data mimicking the spatial variability of satellite based CO_2 measurements.

Lots of data, decent spatial coverage, but more “missing” than “observed”.

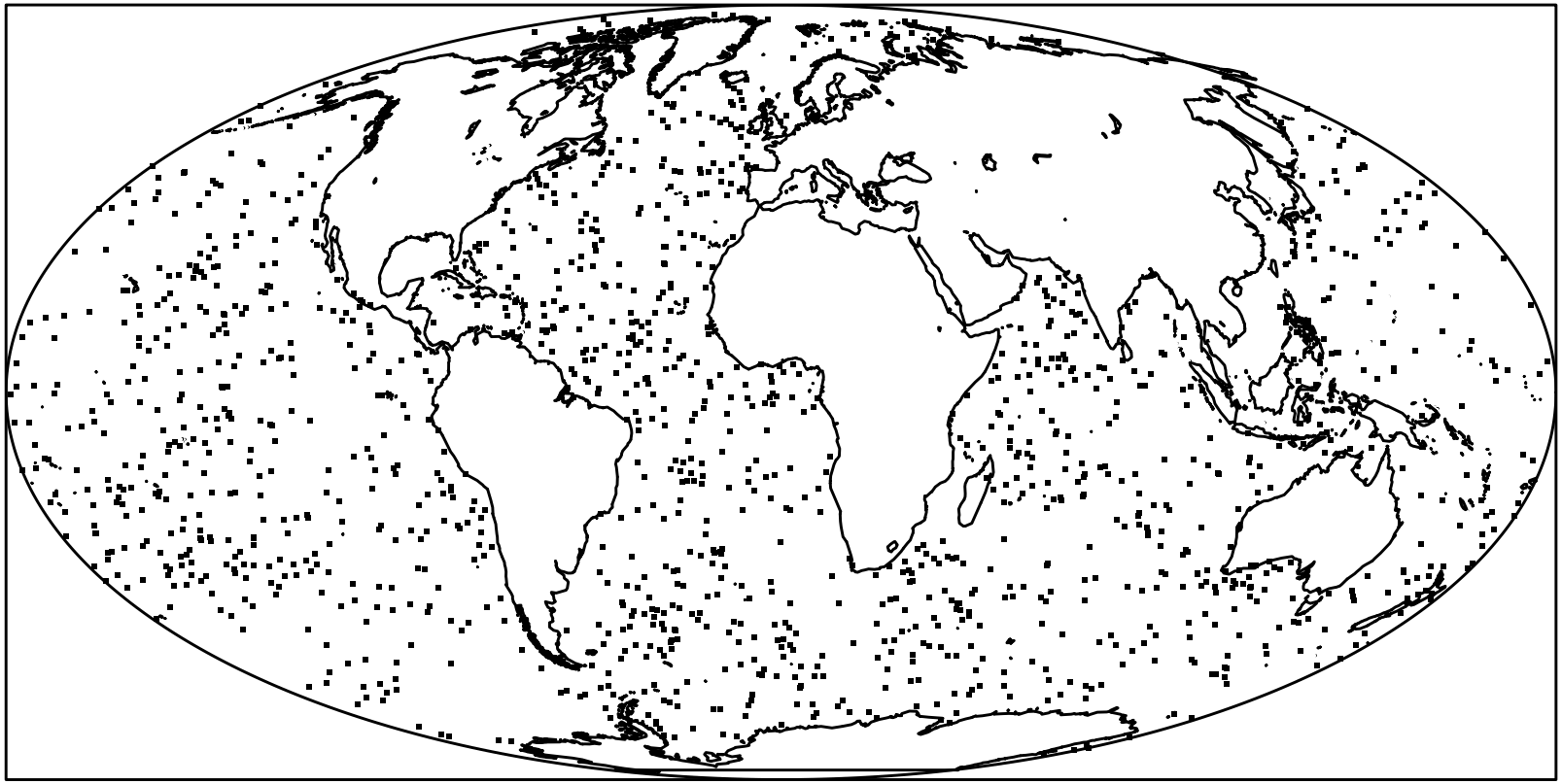
Sparse spatial coverage of temperature measurements



Total observations: ≈ 20,000,000 from daily timeseries over 160 years

Spatial information: 350 observations

Point process data are very weakly informative



Big data, information, and stochastic models

Summary

- Space-time data is nearly always sparse and spatially uneven
- Nearly *all* potential observables are “missing”
- Fortunately, “missingness” is a problem only for specific estimation algorithms, not for statistical analysis as such
- Hierarchical/generative/Bayesian/regularization models provide a general and useful framework for analysis of multi data source data sets

Challenges:

- Even larger output than input
- Quantifying uncertainty is harder than reasonable point estimates
- Can we use physical world knowledge?
- Representing multi-scale stochastic dependence
- ...in a numerics friendly manner
- Hierarchical models are general, but MCMC is very slow
- We *can* use a lot more direct computation

Brief introduction to spatial modelling

Spatial statistics framework

- Spatial domain Ω, or space-time domain $\Omega \times \mathbb{T}$, $\mathbb{T} \subset \mathbb{R}$.
- Random field $u(\boldsymbol{s})$, $\boldsymbol{s} \in \Omega$, or $u(\boldsymbol{s}, t)$, $(\boldsymbol{s}, t) \in \Omega \times \mathbb{T}$.
- Observations y_i. Usually $y_i = u(\boldsymbol{s}_i) + \epsilon_i$ (simple georeferenced data in geostatistics) or more generally $y_i \sim$ GLMM, with $u(\cdot)$ as a structured random effect.
- We'll focus on Bayesian hierarchical models in the form of spatial latent Gaussian models.

Two basic model and method components

- We need stochastic models for $u(\cdot)$.
- We need computationally efficient (Bayesian) inference methods for the posterior distribution of $u(\cdot)$ given data $\boldsymbol{y}$.

Covariance functions and stochastic PDEs

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

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

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

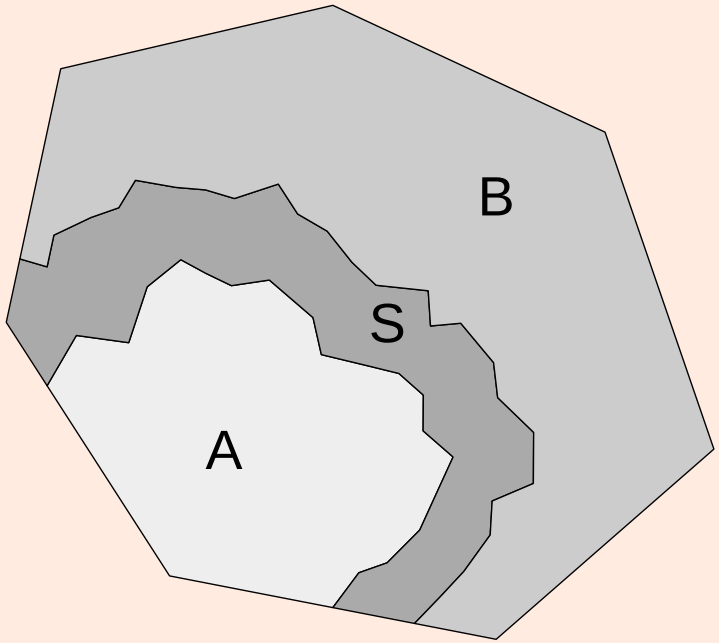
$$\left(\kappa^2 - \nabla \cdot \nabla\right)^{\alpha/2} u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \alpha = \nu + d/2$$

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

Continuous and discrete Markov properties

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.
(Rozanov, 1977)

Discrete representations ($\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$):

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

Continuous domain Markov approximations

Continuous Markovian spatial models (Lindgren et al, 2011)

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

Basis weights: $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1})$, 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: $\boldsymbol{Q} = \kappa^4 \boldsymbol{C} + 2\kappa^2 \boldsymbol{G} + \boldsymbol{G}_2$ $\quad (\kappa^4 + 2\kappa^2 |\boldsymbol{\omega}|^2 + |\boldsymbol{\omega}|^4)$

Conditional distribution in a Gaussian model

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

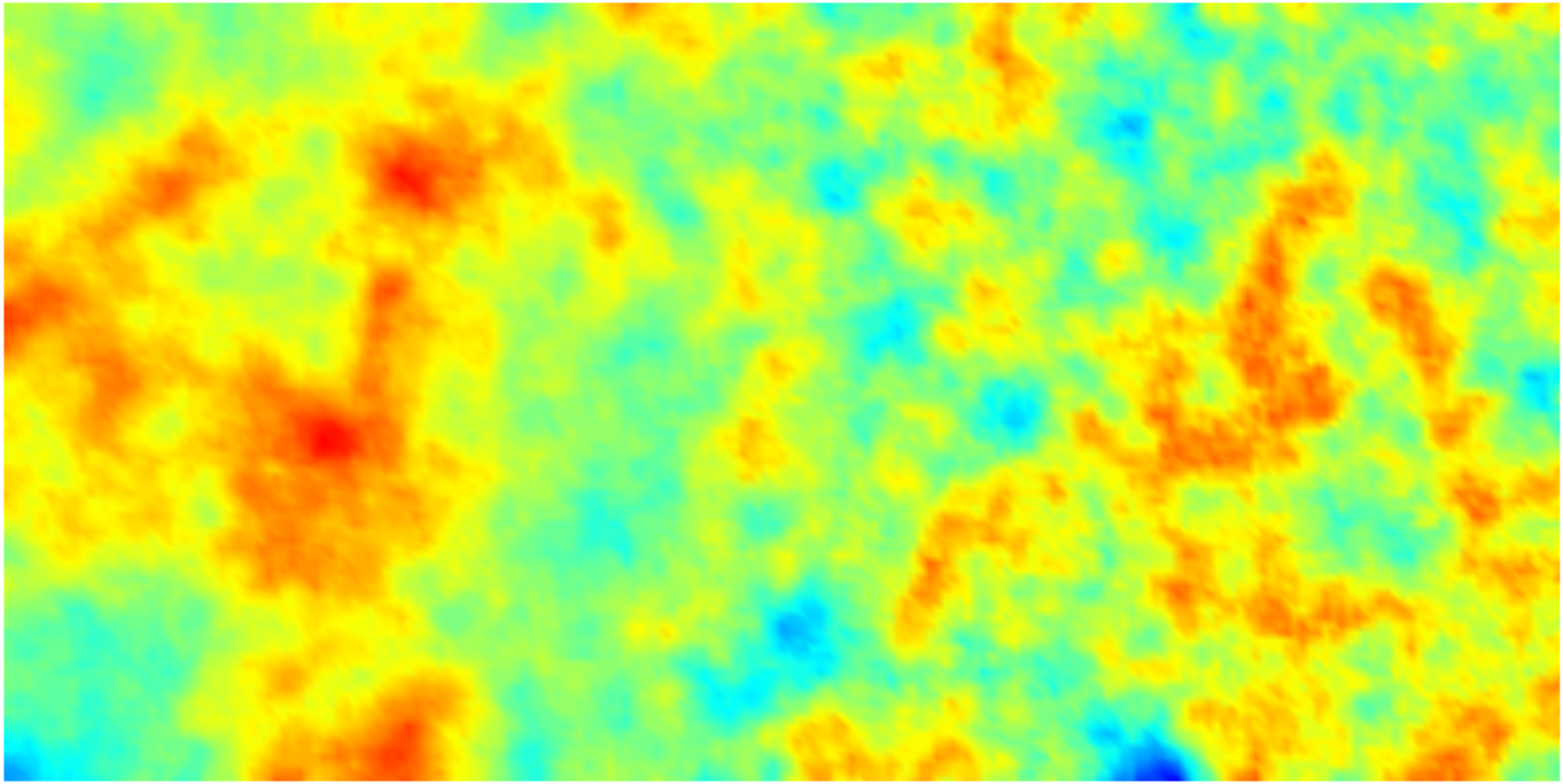
$$\boldsymbol{u}|\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_{u|y}, \boldsymbol{Q}_{u|y}^{-1})$$

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

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

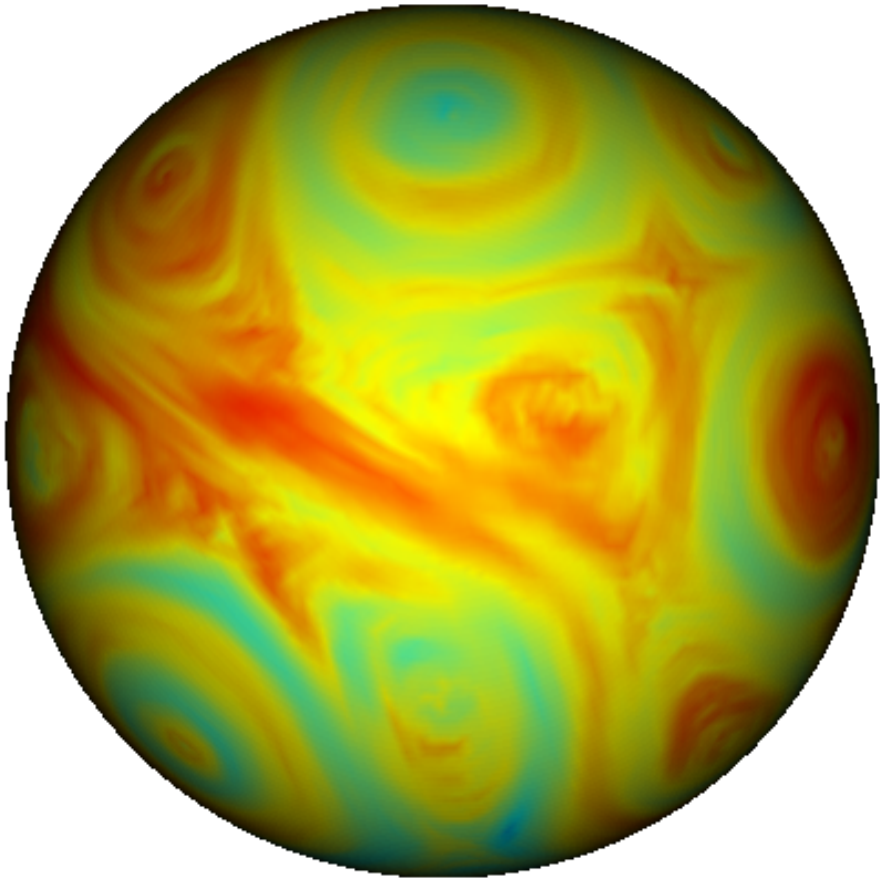
We've translated the spatial inference problem into sparse numerical linear algebra similar to finite element PDE solvers

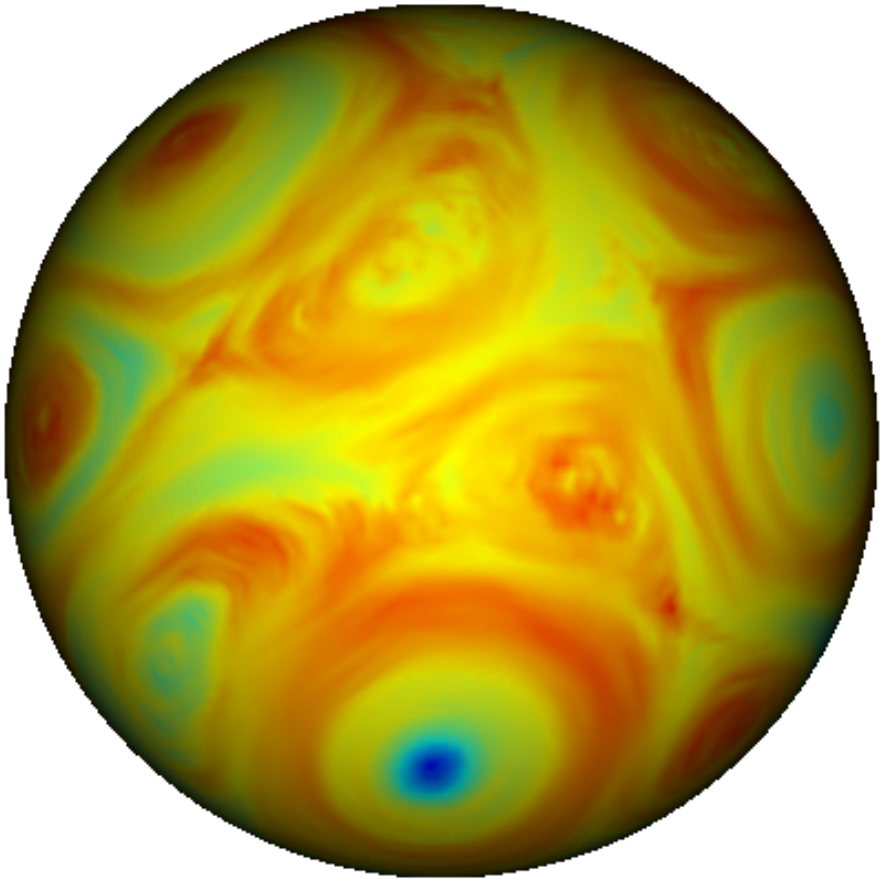
Non-stationary field



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

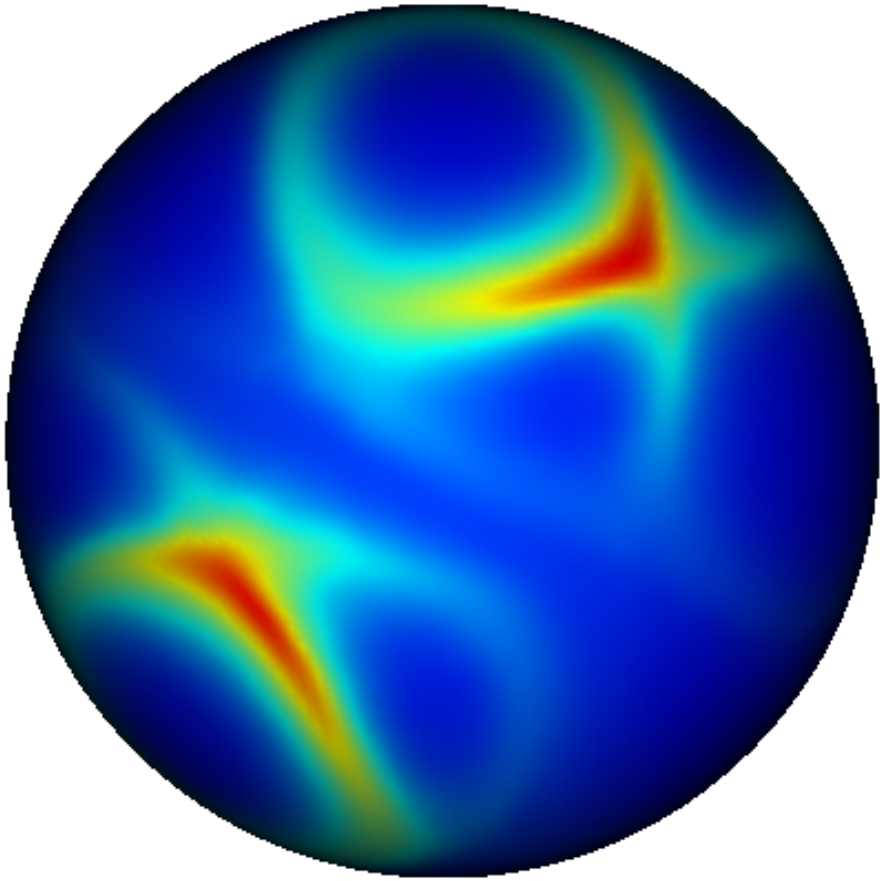
Anisotropic field on a globe via vector parameter field

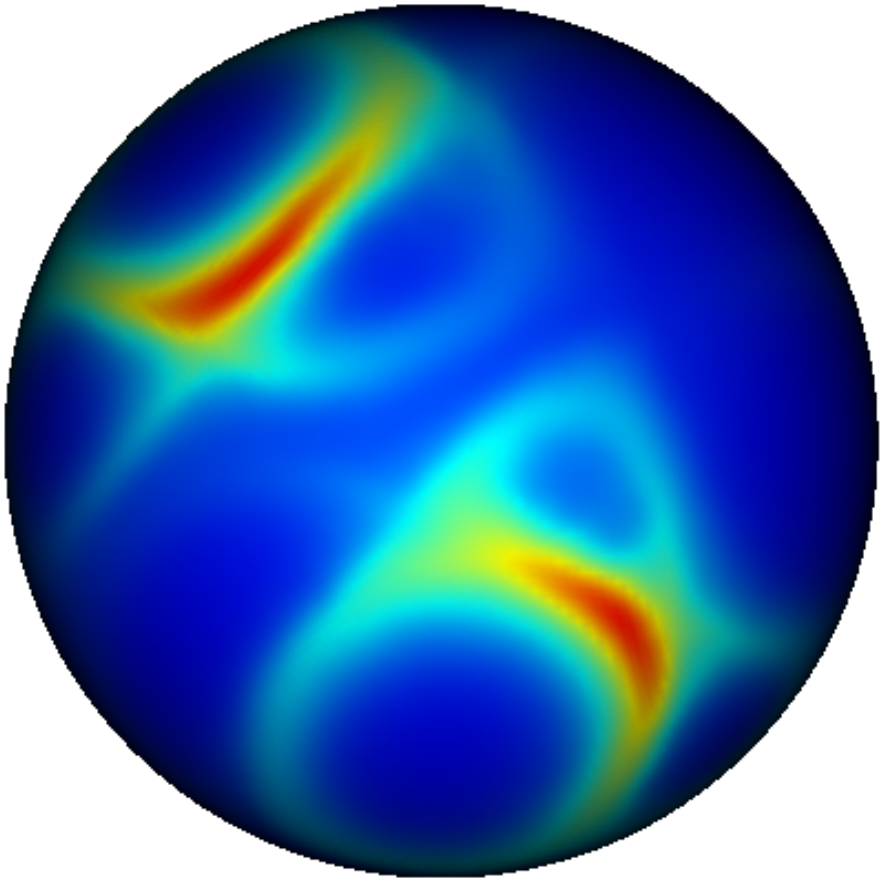




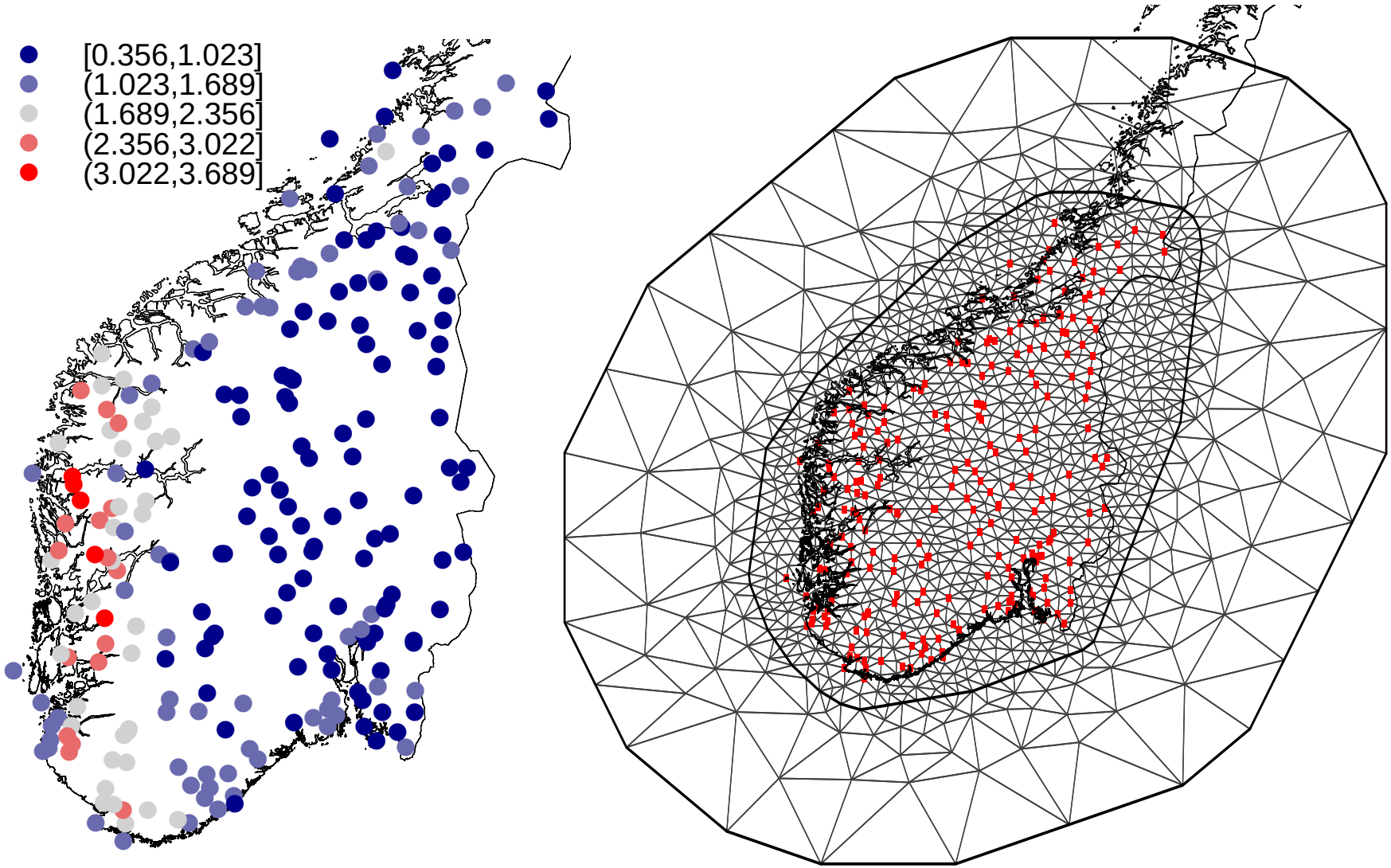
$$(\kappa(\boldsymbol{s})^2 - \nabla \cdot \boldsymbol{H}(\boldsymbol{s})\nabla)u(\boldsymbol{s}) = \kappa(\boldsymbol{s})\mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \Omega$$

Four covariance functions





Example: Precipitation (Ingebrigtsen et al., 2013)



Non-stationary precision construction

Finite element construction of basis weight precision

Non-stationary SPDE:

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

The SPDE parameters are constructed via spatial covariates:

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

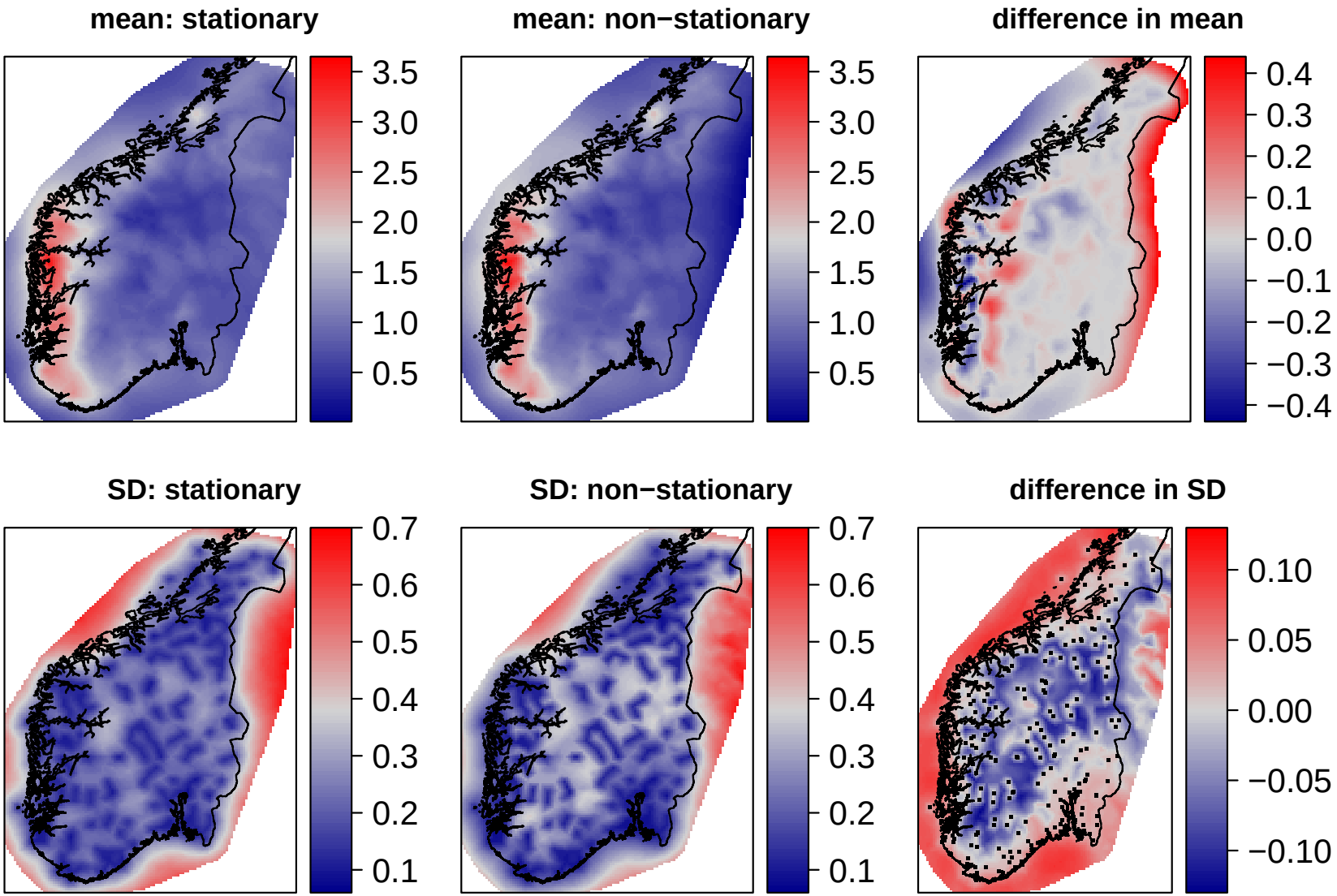
Finite element calculations give

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

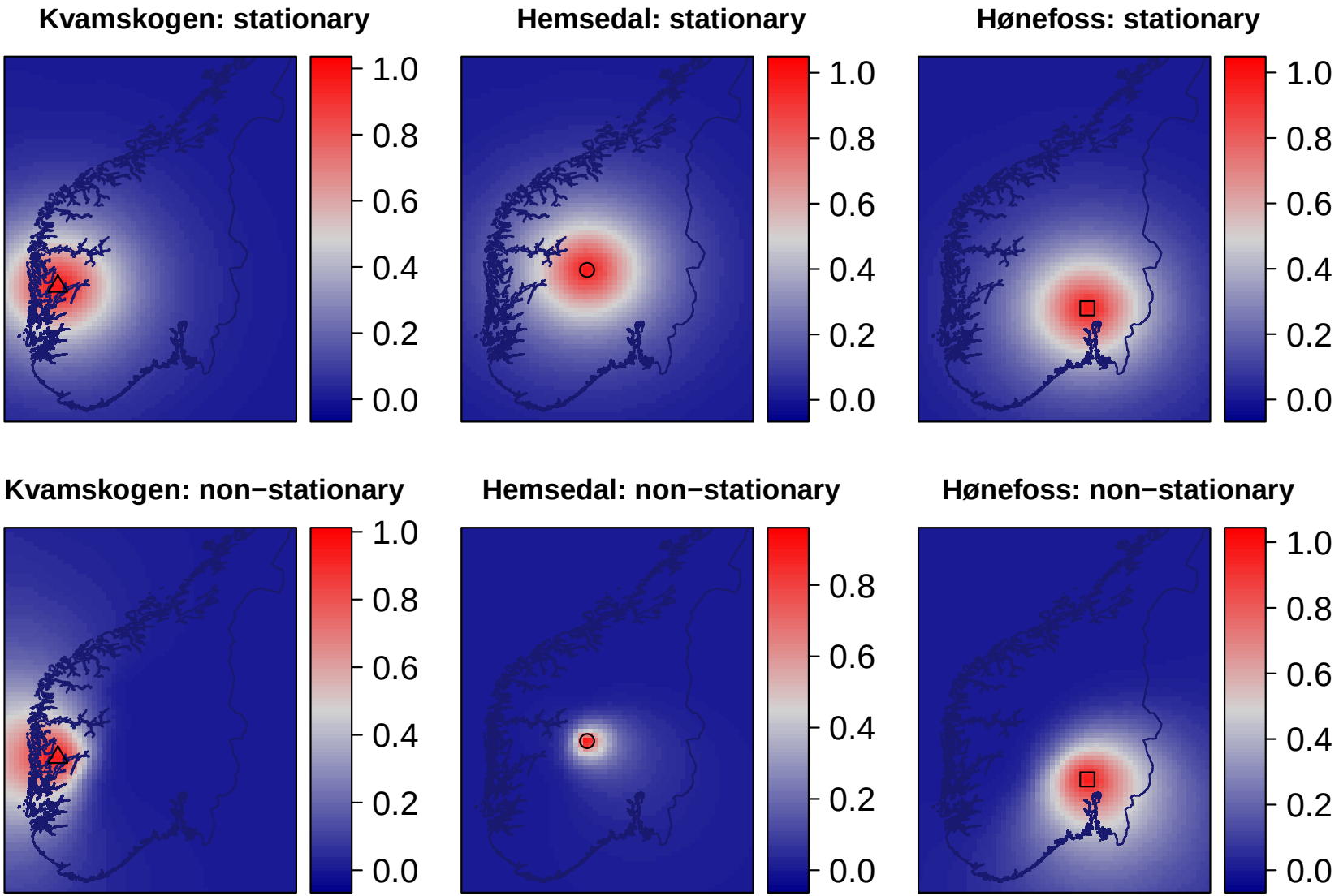
$$C_{ii} = \int \psi_i(\boldsymbol{s})\, d\boldsymbol{s}, \quad G_{ij} = \int \nabla\psi_i(\boldsymbol{s}) \cdot \nabla\psi_j(\boldsymbol{s})\, d\boldsymbol{s}$$

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

Results for stationary and non-stationary models



Correlations for stationary and non-stationary models



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:

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

The likelihood can be approximated numerically, e.g.

$$\int_\Omega \lambda(\boldsymbol{s})\, d\boldsymbol{s} \approx \sum_{j=1}^{N} \lambda(\boldsymbol{s}_j) w_j$$

Laplace approximations

Quadratic posterior log-likelihood approximation

$$p(\boldsymbol{u} \mid \boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\mu}_u, \boldsymbol{Q}_u^{-1}), \quad \boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\theta} \sim p(\boldsymbol{y} \mid \boldsymbol{u})$$

$$p_G(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta}) \sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{Q}}^{-1})$$

$$\mathbf{0} = \nabla_{\boldsymbol{u}} \left\{ \ln p(\boldsymbol{u} \mid \boldsymbol{\theta}) + \ln p(\boldsymbol{y} \mid \boldsymbol{u}) \right\}\big|_{\boldsymbol{u}=\widetilde{\boldsymbol{\mu}}}$$

$$\widetilde{\boldsymbol{Q}} = \boldsymbol{Q}_u - \nabla_{\boldsymbol{u}}^2 \ln p(\boldsymbol{y} \mid \boldsymbol{u})\big|_{\boldsymbol{u}=\widetilde{\boldsymbol{\mu}}}$$

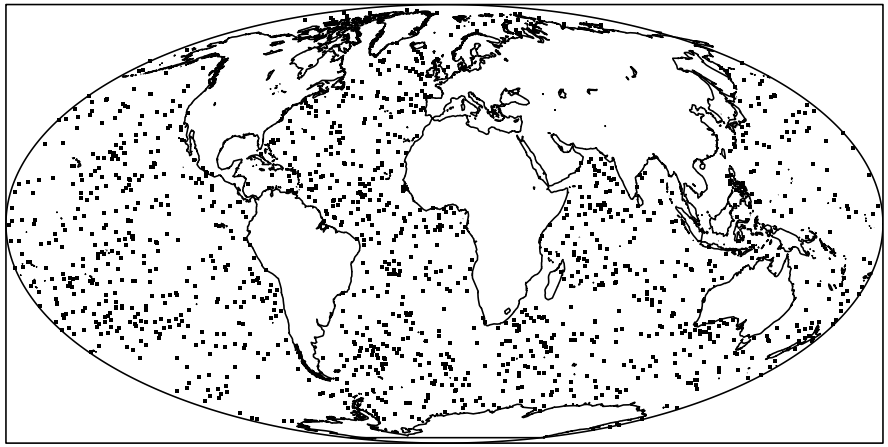
Direct Bayesian inference with INLA (`r-inla.org`)

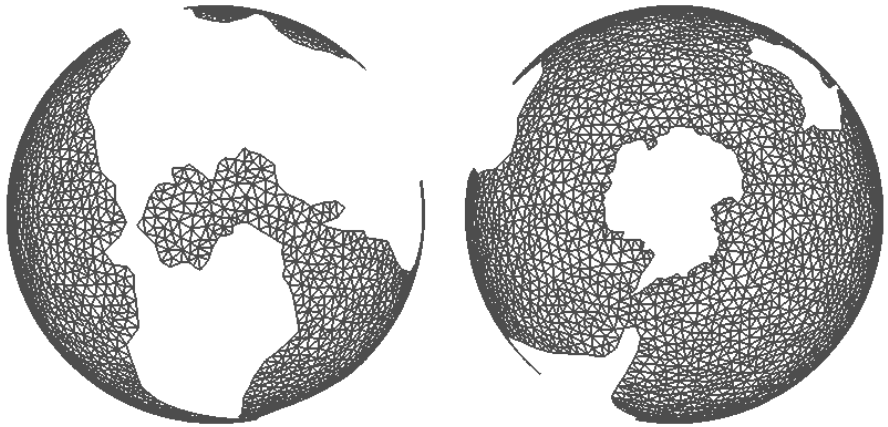
$$\widetilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \left. \frac{p(\boldsymbol{\theta}) p(\boldsymbol{u} \mid \boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\theta})}{p_G(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta})} \right|_{\boldsymbol{u}=\widetilde{\boldsymbol{\mu}}}$$

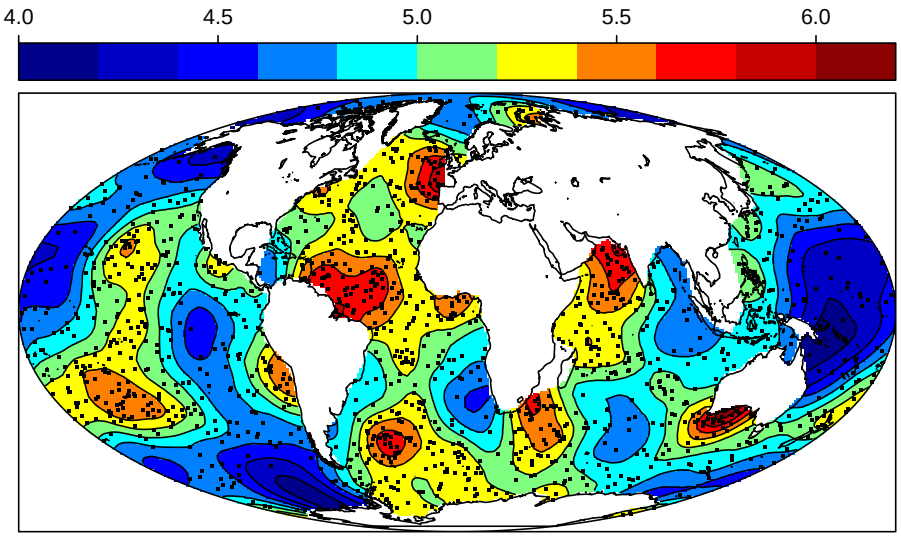
$$\widetilde{p}(\boldsymbol{u}_i \mid \boldsymbol{y}) \propto \int p_{GG}(\boldsymbol{u}_i \mid \boldsymbol{y}, \boldsymbol{\theta}) \widetilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) \, d\boldsymbol{\theta}$$

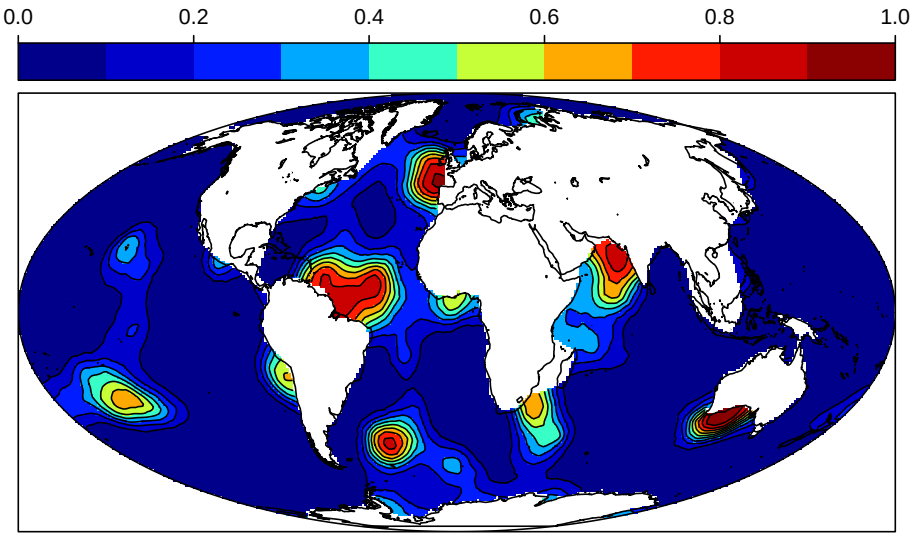
For a large model class this approach is much faster, as well as more accurate, than MCMC

SPDE based inference for point process data

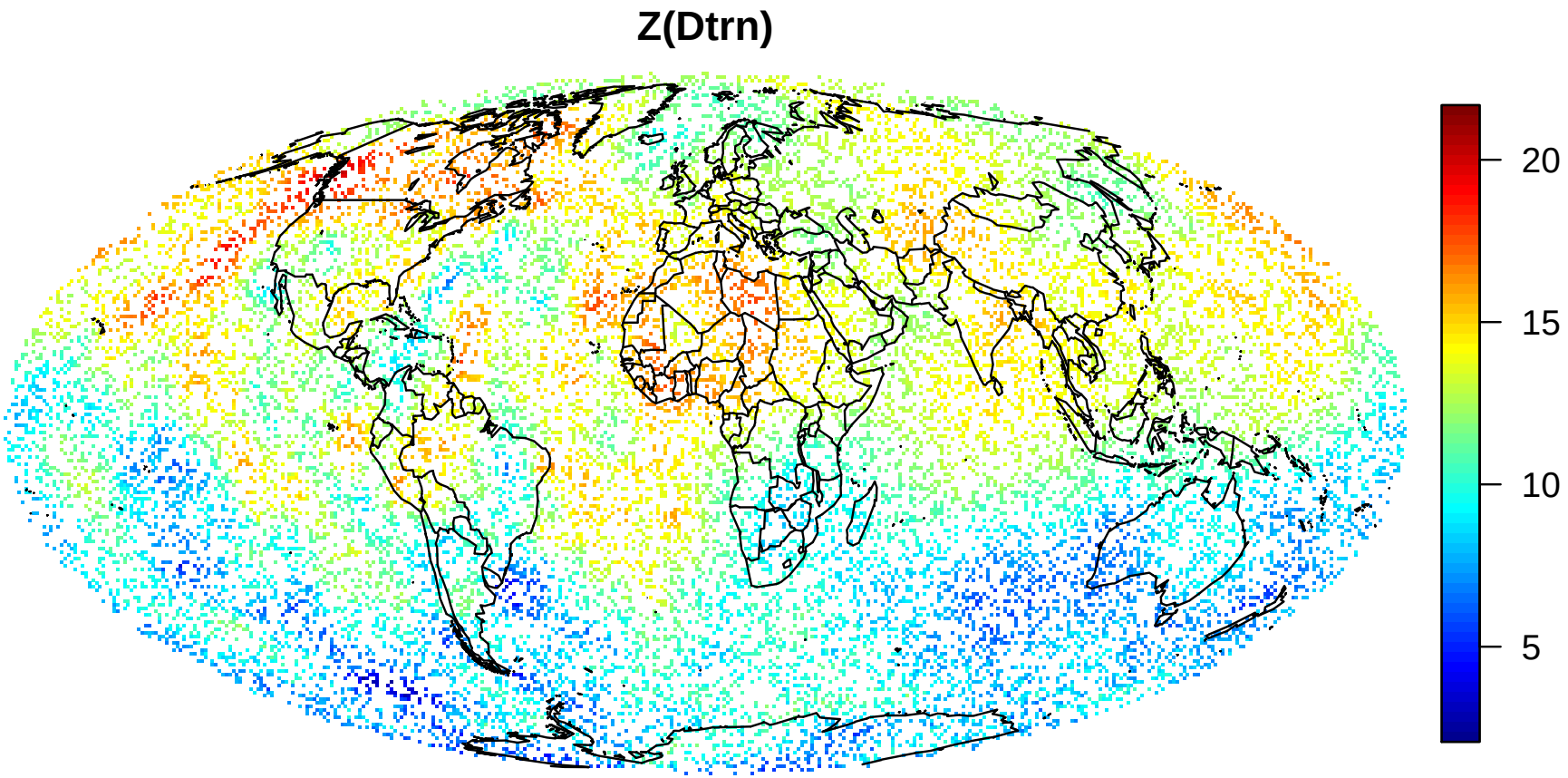




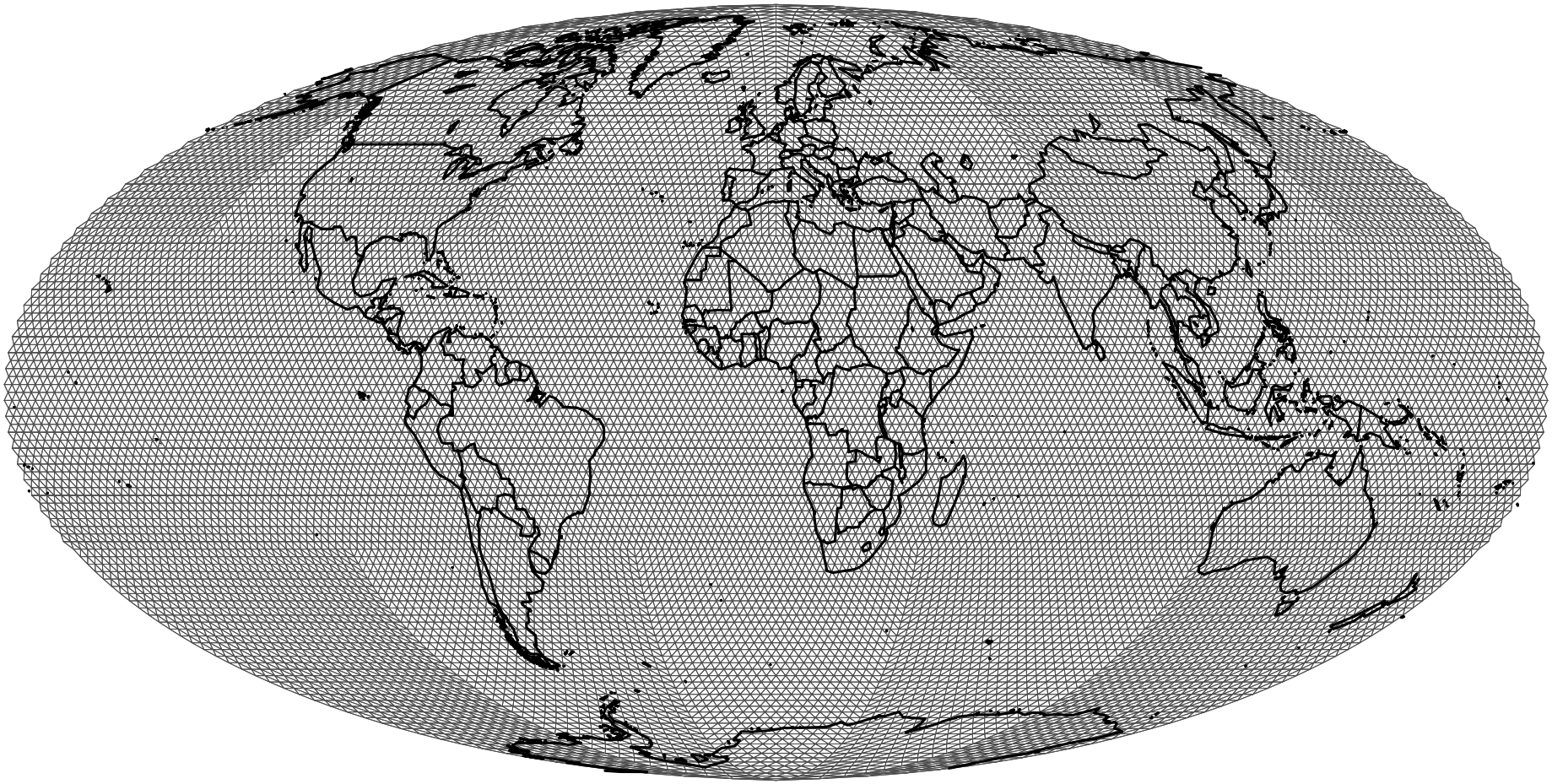




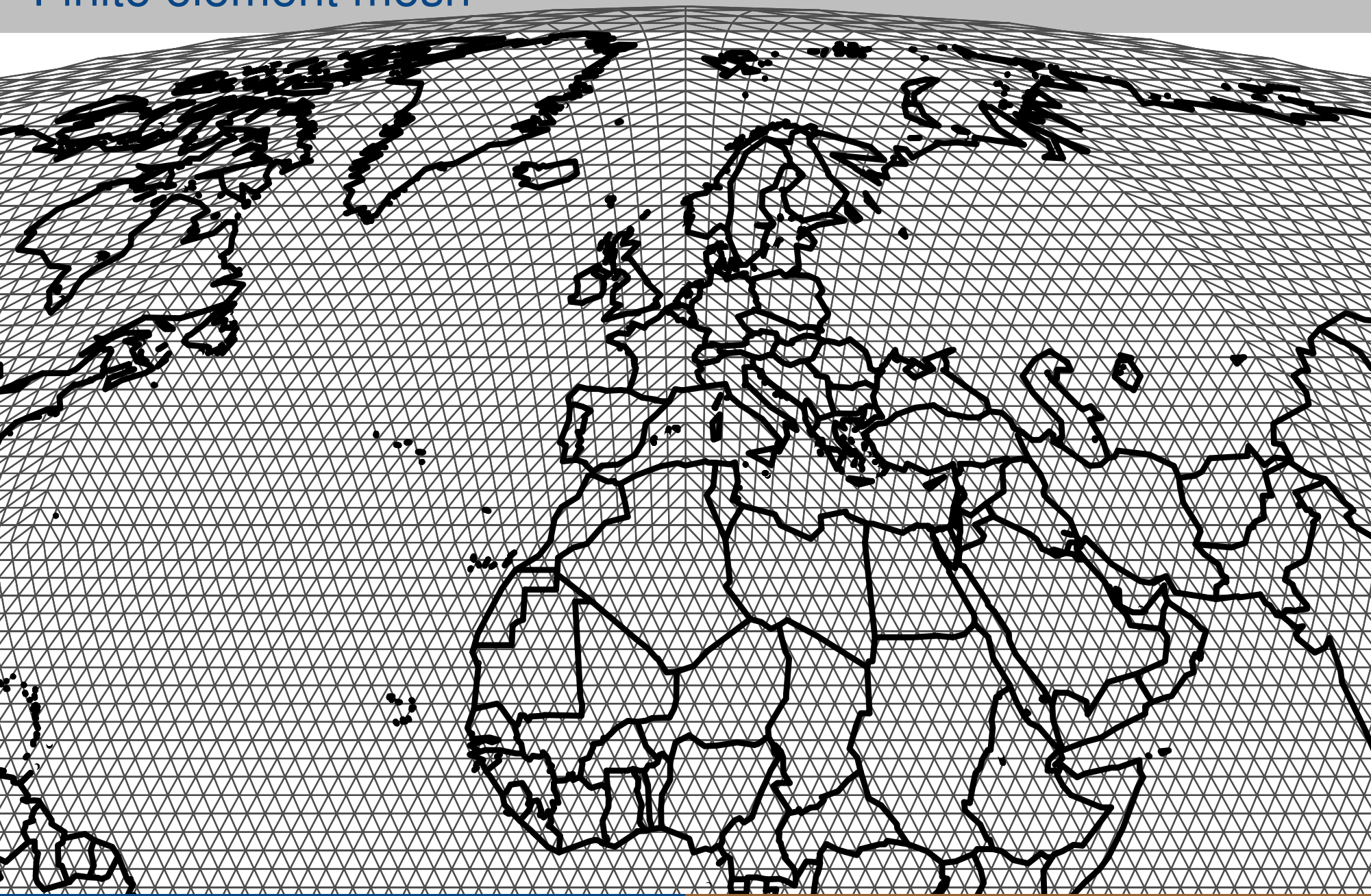
Synthetic CO2 data



Finite element mesh

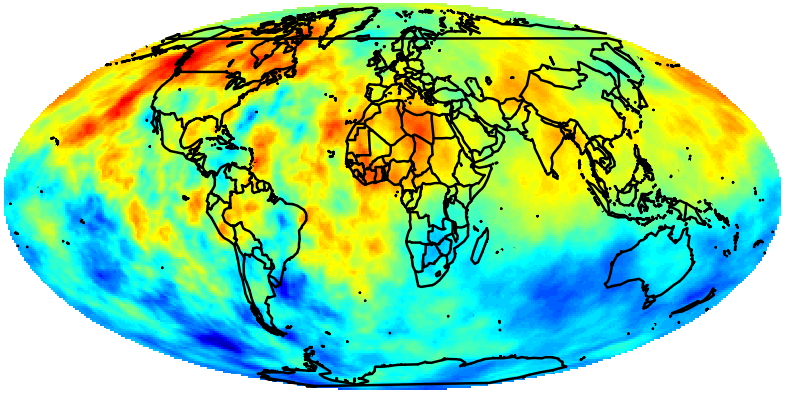


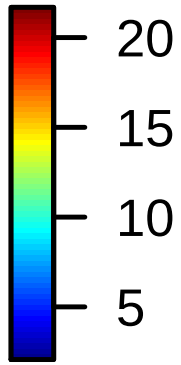
Finite element mesh



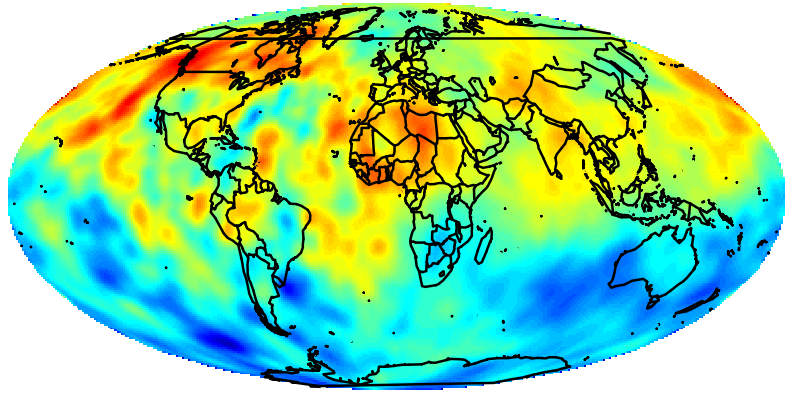
Estimates of CO2 from synthetic data

Ytrue

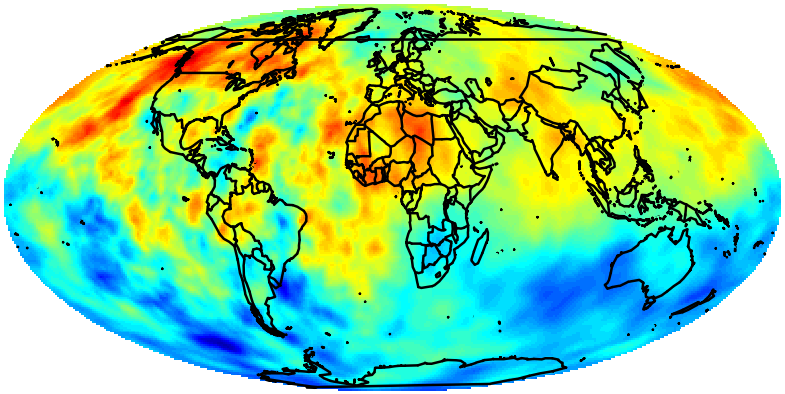


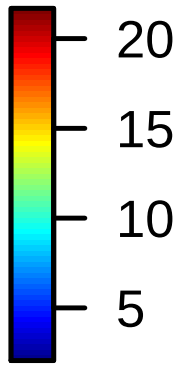


YLKR

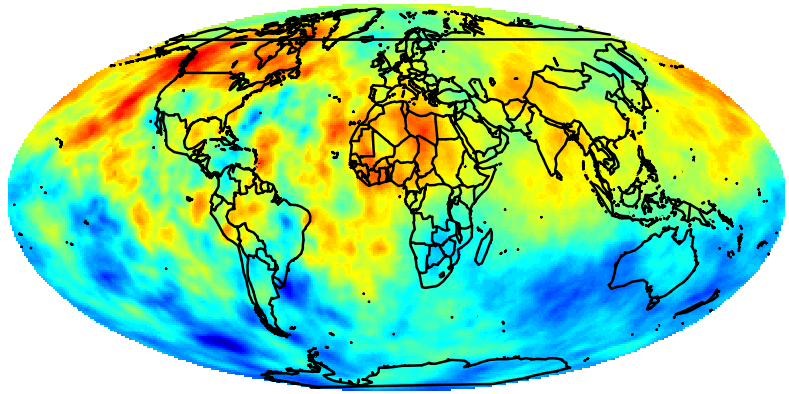


Y0

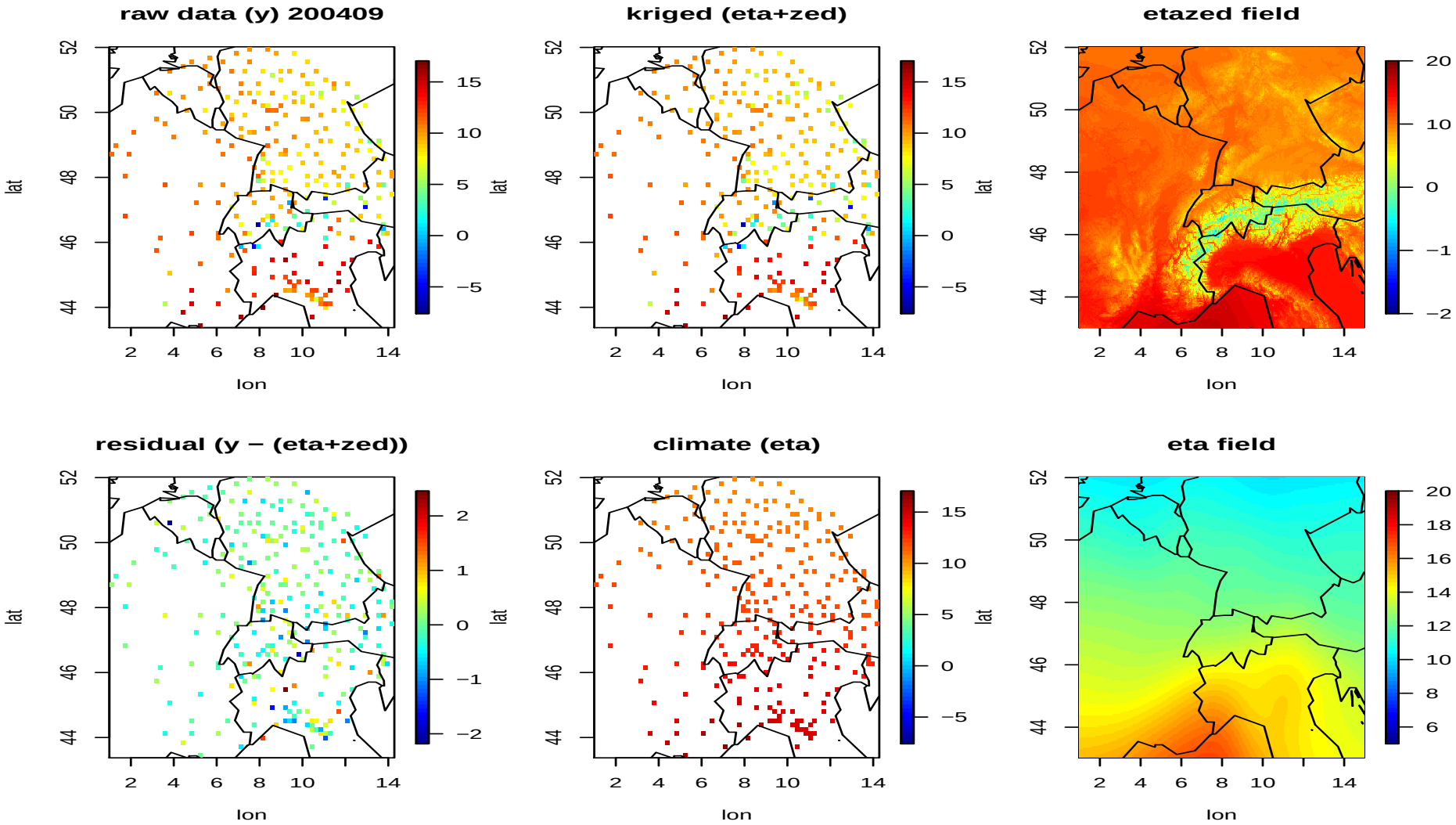




Y1

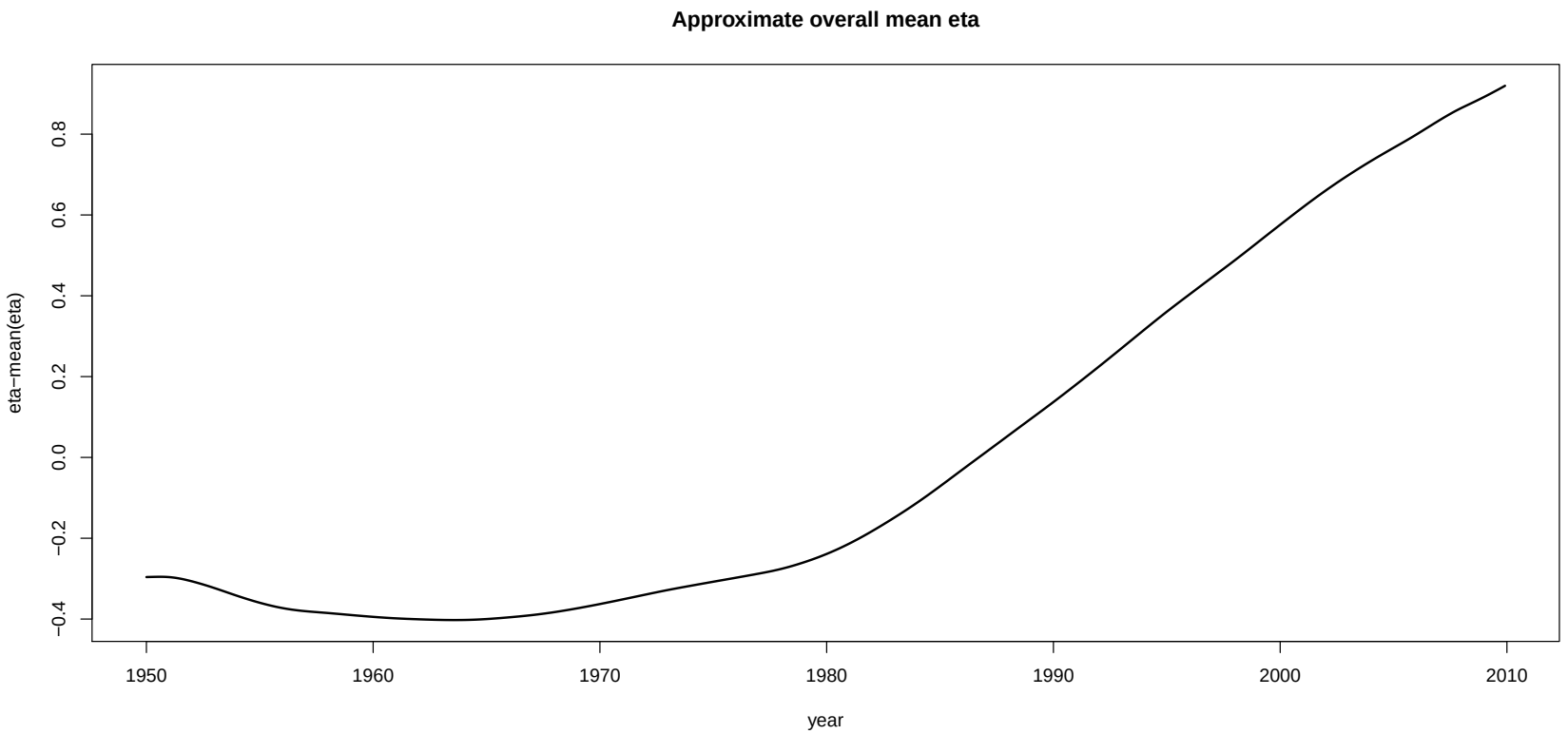


Temperature estimation with geographical covariates and stochastic fields



Total observations: ≈ 90 000 000 from daily timeseries over 160 years

Spatial average climate estimate



Model:

weather $=$ elevation effect $+$ slow climate process $+$ seasonal effects $+$ fast weather process $+$ observation noise

The computational work-horse

Cholesky decomposition (Cholesky, 1924)

$$\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^\top, \quad \boldsymbol{L} \text{ lower triangular}$$

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

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

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)

The end!