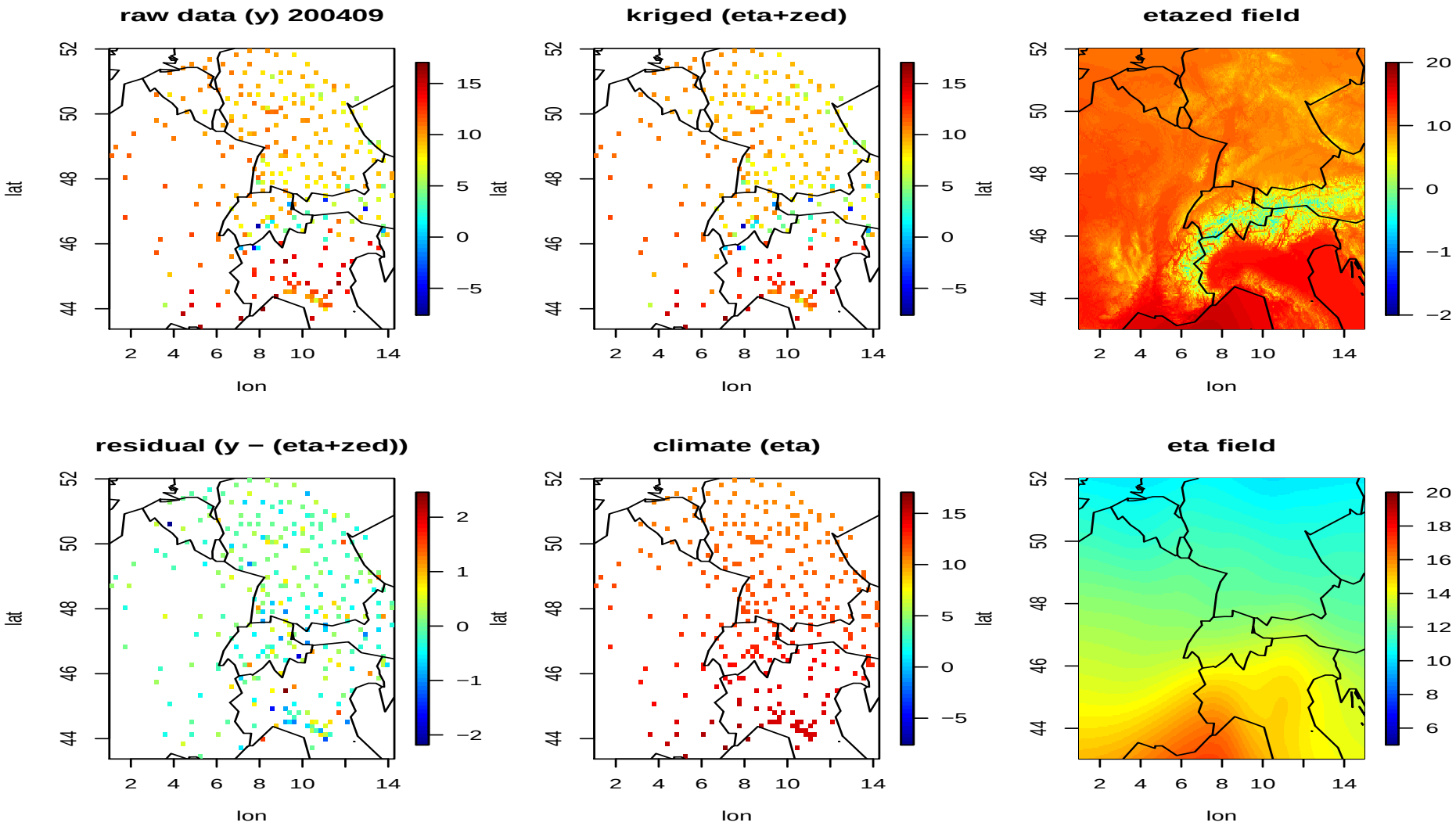
Towards realistic stochastic modelling of global *daily* temperatures

Finn Lindgren

Big Data in Environmental Science
University of British Columbia, PIMS
Vancouver, 12 May 2015

Sparse spatial coverage of temperature measurements



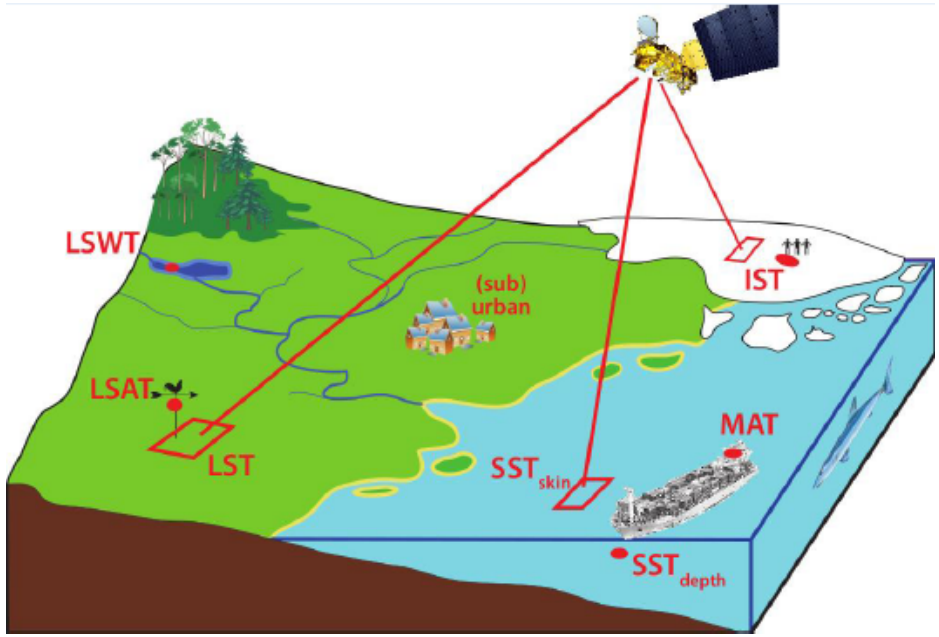
300 regional stations: ≈ 20,000,000 from daily timeseries over 160 years

Full data: 70000 stations, several satellites, many ships, some lakes

EUSTACE (*EU Surface Temperatures for All Corners of Earth*)

EUSTACE has received funding from the European Union's Horizon 2020 Programme for Research and Innovation, under Grant Agreement no 640171

EUSTACE will give publicly available daily estimates of surface air temperature since 1850 across the globe for the first time by combining surface and satellite data using novel statistical techniques.



Nick A. Rayner, Renate Auchmann, Janette Bessembinder, Stefan Brönnimann, Yuri Brugnara, Laura Carrea, Darren Ghent, Elizabeth Good, Katie Herring, Jacob Høyer, John Kennedy, Albert Klein Tank, Finn Lindgren, Colin Morice, Chris Merchant, John Remedios, Ag Stephens and Rasmus Tonboe

contact: nick.rayner@metoffice.gov.uk

EUSTACE data and modelling workflow

- WP1: *Translate sensor data into local estimates of air temperatures*
  Relationships vary over land, ocean, ice, and lakes, and with season
  The estimation errors are spatially correlated
  Daily station data shows timeseries breaks in mean and distribution
- WP2: *Spatio-temporal blending of all data sources*
  Two approaches: “Advanced traditional” and “Ambitious”
  Point estimates of air temperature and uncertainty, and a random sample/ensemble from the posterior distribution
- WP3: *Validation of calibration models and data products*

Modelling challenges:

- Each data sources generates daily values of one (or more) of T_{max}, T_{min}, T_{avg}, and T_{range} at irregular locations
- T_{max} and T_{min} are strongly non-Gaussian and dependent, but T_{avg} and T_{range} are less dependent.
- After compensating for the seasonal cycle, T_{avg} is close to Gaussian, but T_{range} is non-Gaussian.

Preliminary model framework

- Build a priori independent multi-component models for T_{avg} and T_{range}, in space ($\mathbf{s}$) and time (year t and time within year τ):

$$T_{\text{avg}}(\mathbf{s}, t, \tau) = \text{seasonal}_{\text{avg}}(\mathbf{s}, \tau) + \text{longterm}_{\text{avg}}(\mathbf{s}, t, \tau) + \text{shortterm}_{\text{avg}}(\mathbf{s}, t, \tau)$$

  A transformation or copula model is needed for the range, non-linear $h(\cdot)$:

$$\log\left(\text{scaling}_{\text{range}}(\mathbf{s}, t, \tau)\right) = \text{seasonal}_{\text{range}}(\mathbf{s}, \tau) + \text{longterm}_{\text{range}}(\mathbf{s}, t, \tau)$$

$$T_{\text{range}}(\mathbf{s}, t, \tau) = \text{scaling}_{\text{range}}(\mathbf{s}, t, \tau) \cdot h\left(\text{shortterm}_{\text{range}}(\mathbf{s}, t, \tau)\right)$$

  Each component is a Gaussian process with different spatial and temporal covariance properties.

- For observations of daily maxima and minima, define

$$T_{\text{min}} = T_{\text{avg}} - T_{\text{range}}/2$$
$$T_{\text{max}} = T_{\text{avg}} + T_{\text{range}}/2$$

Multiple data sources, all with their own issues

- Ground station air temperatures, with temporally persistent systematic deviations, e.g.
  $y_i^{\text{air,max}}(t) = T_{\text{max}}(\boldsymbol{s}_i, t) + \sum_{k=1}^{K_i} H_k(t)\beta_{ik} + \epsilon_i(t)$,
  where $H_k(\cdot)$ typically are step functions around time series breakpoints.
- Ship measurements and buoys. Similar to ground stations, but moving around in space! And they measure water temperature as well. Sometimes with buckets.
- Geostationary and polar orbiting satellite measurements of land surface, ocean surface, and ice sheet surface temperatures. Different measurement footprints, spatially correlated errors, complicated links to air temperatures.
- Lake temperatures. Water temperatures linked with spatio-temporal averages of past air temperatures.

WP1 of EUSTACE will build calibration models for each data source, with uncertainties tracked into pseudo-observation models for air temperature, which is then fed into a unified Bayesian spatio-temporal estimate in WP2, via

$$p(T_{\text{avg}}, T_{\text{range}}, \boldsymbol{\beta} \mid Y_1, \ldots Y_N) \propto p(T_{\text{avg}}, T_{\text{range}}, \boldsymbol{\beta}) \prod_{k=1}^{N} p(Y_k \mid T_{\text{avg}}, T_{\text{range}}, \boldsymbol{\beta})$$

Laplace approximations for non-linear observations

By building the latent random field models as discretisations of stochastic PDEs collecting all the basis coefficients into a single a priori Gaussian vector, the spatio-temporal inference problem is turned into a sparse numerical algebra problem.
The non-linear transformations lead to a non-linear problem that can be approximated with a Gaussian posterior:

Quadratic posterior log-likelihood approximation

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

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

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

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

The product structure for T_{range} doesn't necessarily generate positive definite precisions away from the mode, but the optimisation can use Fisher scoring, where $\mathsf{E}_{\boldsymbol{y}|\boldsymbol{u}}\left(-\nabla_{\boldsymbol{u}}^2 \ln p(\boldsymbol{y} \mid \boldsymbol{u})\right)$ is positive definite.

Products of transformed processes

Assume that $\boldsymbol{u}$ is a large scale process and $\boldsymbol{v}$ is a small scale process, so that they are statistically identifiable from observations of the form

$$y_i = h_u(u_i) \cdot h_v(v_i) + \epsilon_i, \quad h_u \text{ and } h_v \text{ non-linear transformations.}$$

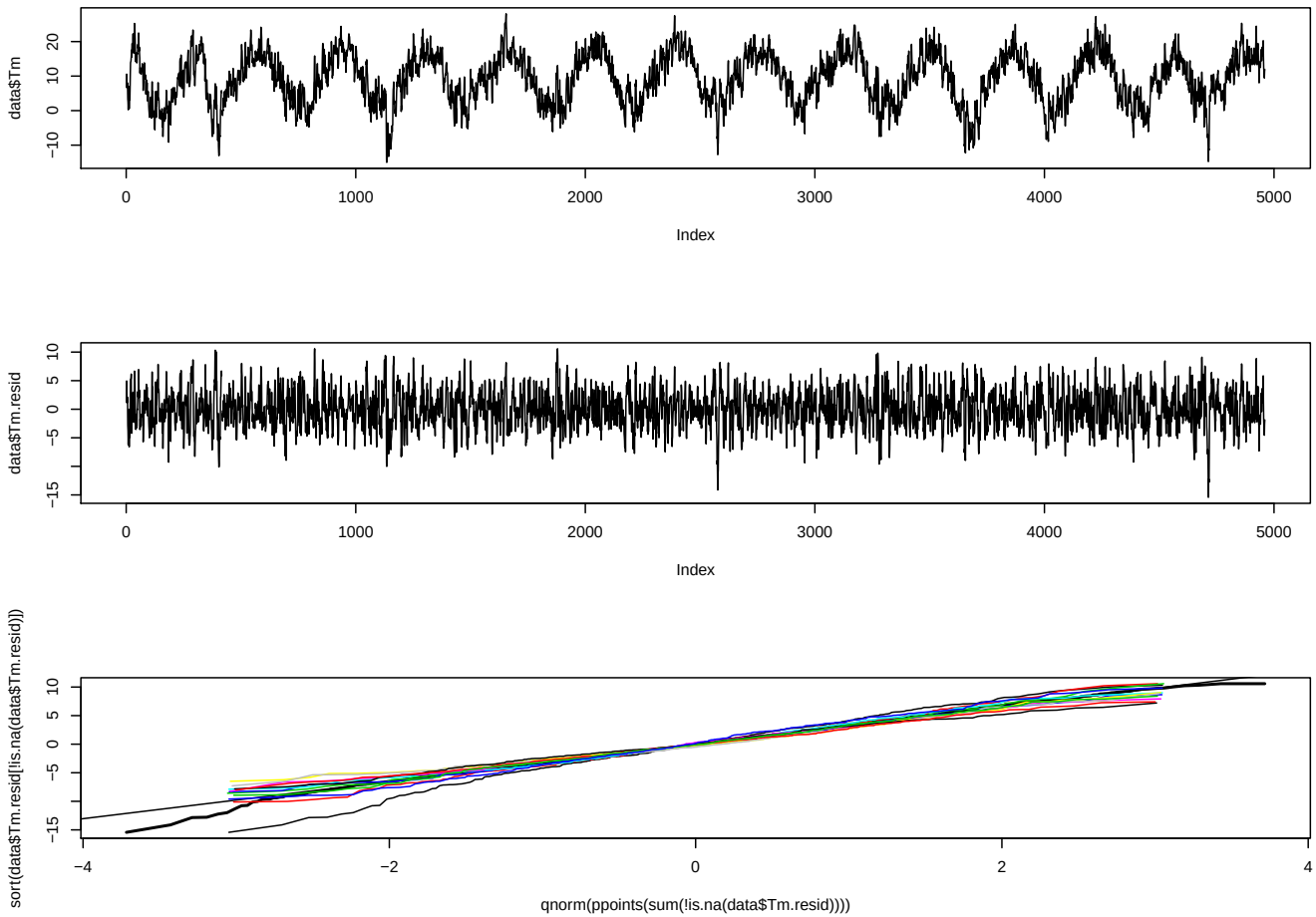
Write $\boldsymbol{h}_u$, $\boldsymbol{h}'_u$, $\boldsymbol{h}''_u$ for the vectors of transformed values and derivatives of h_u at the u_i values, and similarly for $\boldsymbol{v}$. Then

$$\begin{aligned}
C - \log p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{v}) &= \frac{1}{2}(\boldsymbol{y} - \boldsymbol{h}_u \odot \boldsymbol{h}_v)^\top \boldsymbol{Q}_\epsilon (\boldsymbol{y} - \boldsymbol{h}_u \odot \boldsymbol{h}_v) \\
-\frac{\partial}{\partial \boldsymbol{v}} \log p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{v}) &= -\operatorname{diag}(\boldsymbol{h}_u \odot \boldsymbol{h}'_v) \boldsymbol{Q}_\epsilon (\boldsymbol{y} - \boldsymbol{h}_u \odot \boldsymbol{h}_v) \\
-\frac{\partial^2}{\partial \boldsymbol{v}^2} \log p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{v}) &= \operatorname{diag}(\boldsymbol{h}_u \odot \boldsymbol{h}'_v) \boldsymbol{Q}_\epsilon \operatorname{diag}(\boldsymbol{h}_u \odot \boldsymbol{h}'_v) \\
&\quad - \operatorname{diag}(\operatorname{diag}(\boldsymbol{h}_u \odot \boldsymbol{h}''_v) \boldsymbol{Q}_\epsilon (\boldsymbol{y} - \boldsymbol{h}_u \odot \boldsymbol{h}_v))
\end{aligned}$$

and similarly for $\frac{\partial}{\partial \boldsymbol{u}}$, $\frac{\partial^2}{\partial \boldsymbol{u} \partial \boldsymbol{v}}$, and $\frac{\partial^2}{\partial \boldsymbol{u}^2}$. The problematic term in the Hessian involving $\boldsymbol{y}$ disappears in Fisher scoring.

Seasonal effects and shortterm distribution for T_{avg}

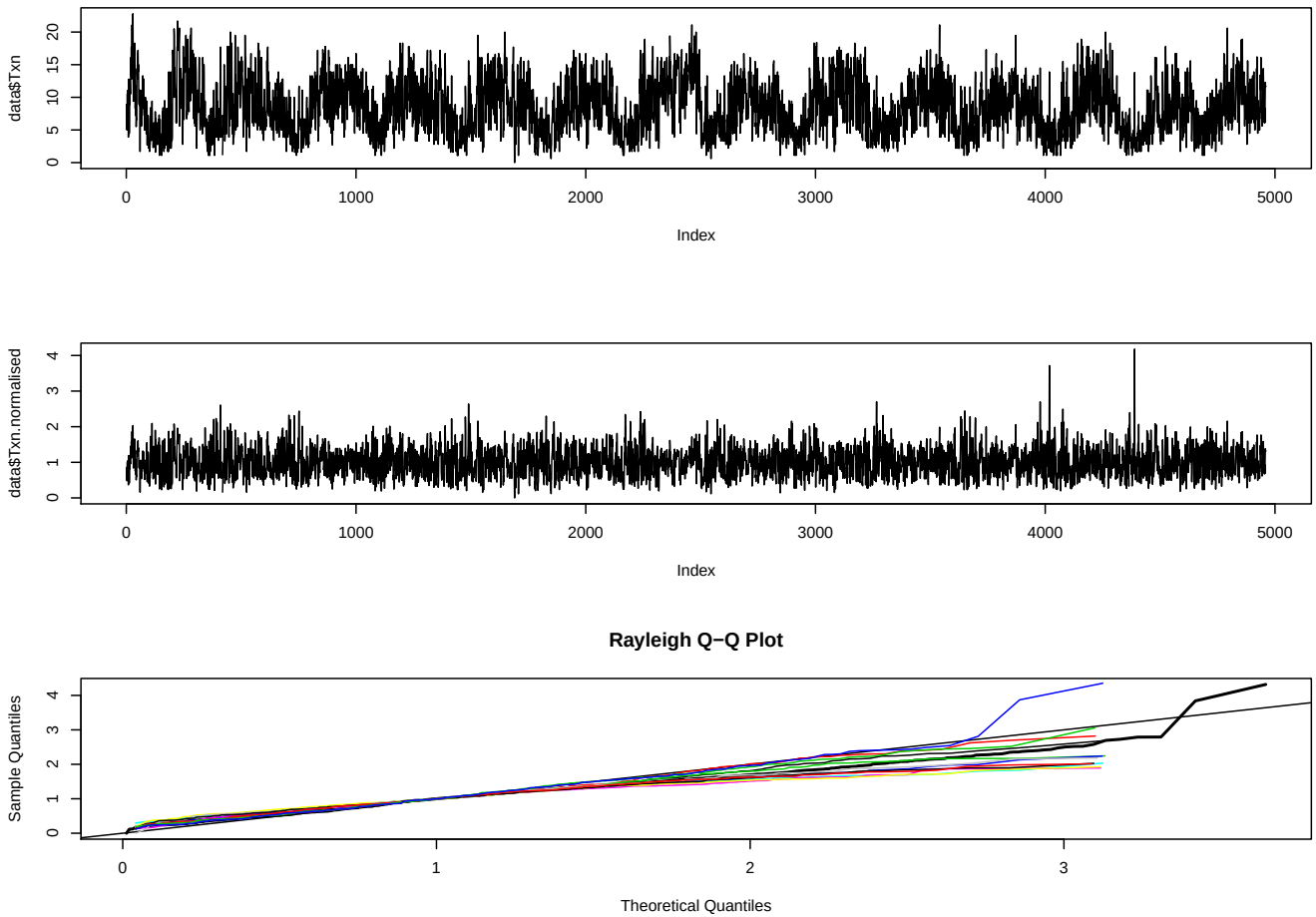
Raw data, estimated short term process, and Normal Q-Q plot for T_{avg}



The residual process is close to Normal, possibly with some seasonal variations.

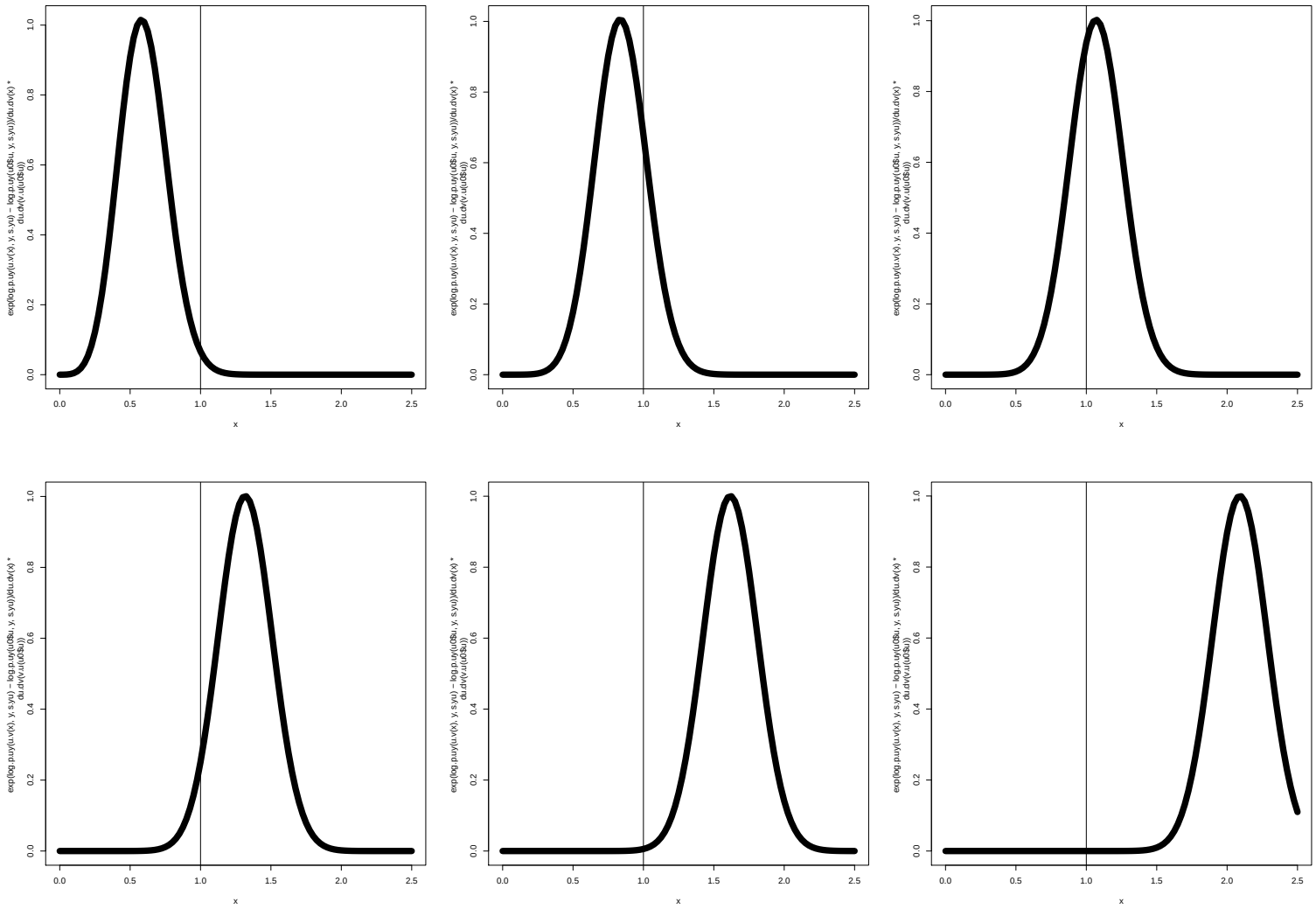
Seasonal effects and shortterm distribution for T_{range}

Raw data, estimated short term process, and Rayleigh Q-Q plot for T_{range}

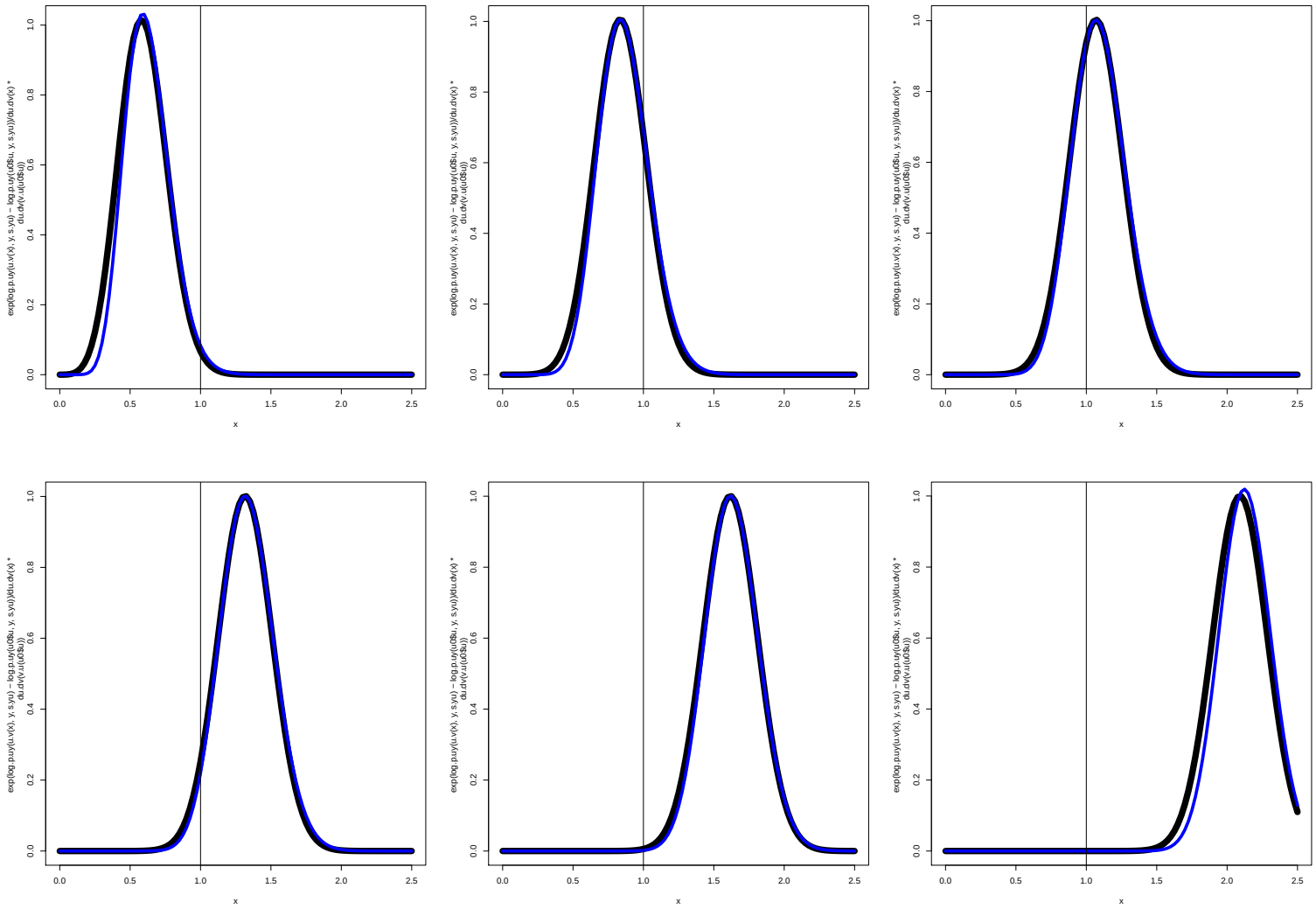


Under some theory, the residual process should have Rayleigh marginal distribution, but the true distribution has lighter tails, with a seasonal pattern.

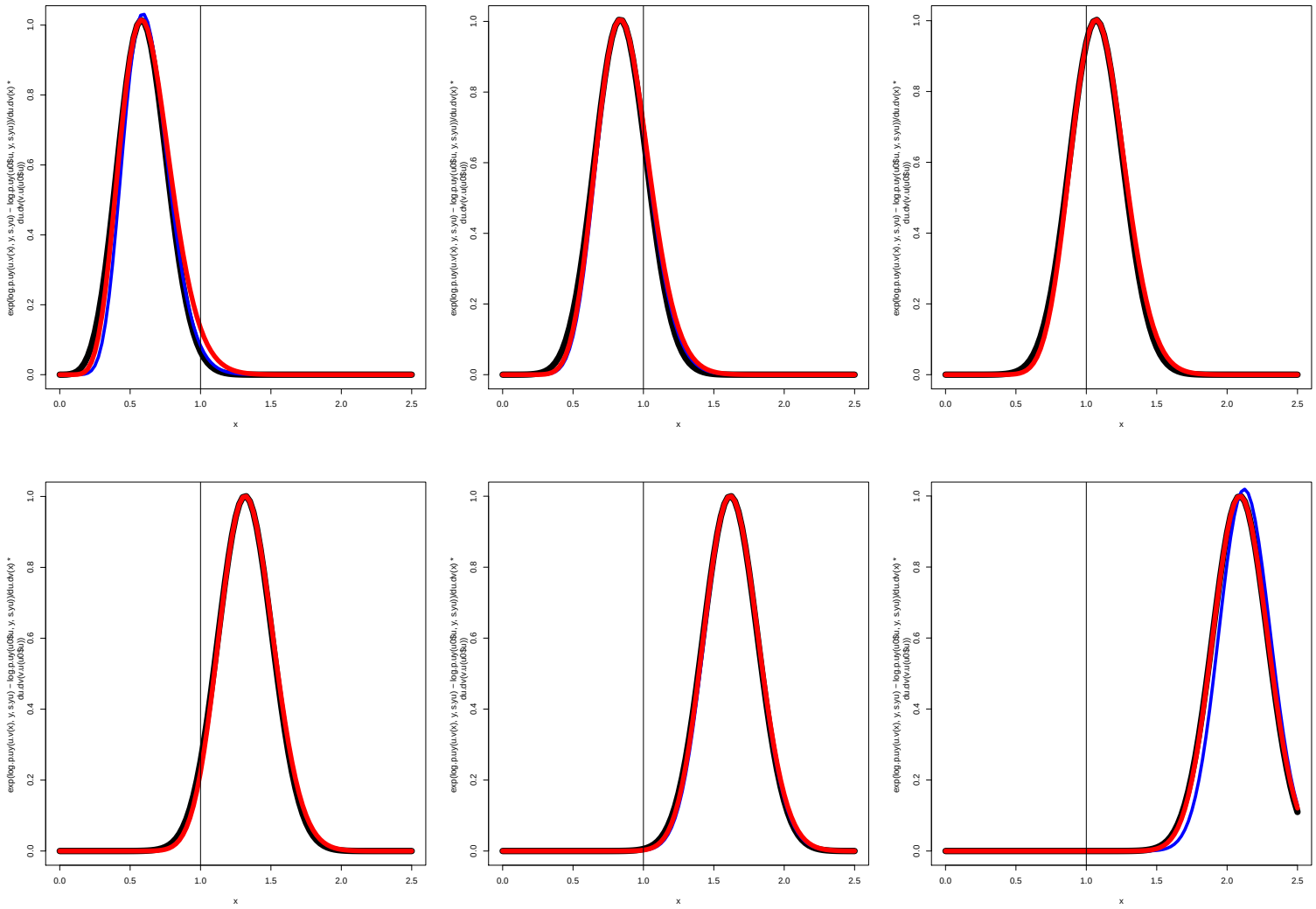
Posterior density approximations with Laplace



Posterior density approximations with Laplace



Posterior density approximations with Laplace



Illustrative modelling and computation principle, revisited

Ignoring the non-linearities, and the seasonal effects:

- A temporally slow, simplified stochastic heat equation (non-separable) for the longterm processes for T_{mean}

$$\frac{\partial}{\partial t} z(\boldsymbol{s}, t) - \gamma_z \nabla \cdot \nabla z(\boldsymbol{s}, t) = \mathcal{E}(\boldsymbol{s}, t)$$

$$(1 - \gamma_{\mathcal{E}} \nabla \cdot \nabla) \mathcal{E}(\boldsymbol{s}, t) = \mathcal{W}_{\mathcal{E}}(\boldsymbol{s}, t)$$

- A temporally quick, spatially non-stationary SPDE/GMRF (separable)
  $\left(\frac{\partial}{\partial t} + \gamma_t\right) \left(\kappa(\boldsymbol{s})^2 - \nabla \cdot \nabla\right) \left(\tau(\boldsymbol{s}) a(\boldsymbol{s}, t)\right) = \mathcal{W}_a(\boldsymbol{s}, t)$
- Direct and linear observations:
  $y_i = a(\boldsymbol{s}_i, t_i) + z(\boldsymbol{s}_i, t_i) + \epsilon_i$, discretised into
  $\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{a} + (\boldsymbol{B} \otimes \boldsymbol{I})\boldsymbol{z}) + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_\epsilon^{-1})$
  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:

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

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

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

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

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

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

GMRF precision for the simplified stochastic heat equation

The precision is a quint-diagonal block matrix, with further structure

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

Ignoring the degenerate aspect of the model, the precision can be pseudo Cholesky factorised as $\boldsymbol{Q}_z = \widetilde{\boldsymbol{L}}_z \widetilde{\boldsymbol{L}}_z^\top$, where

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

Since the kronecker products do not need to be explicitly stored, the pseudo Cholesky factors require very little time and memory, and can be computed even for a very large number of spatial basis functions (up to a million).

Posterior calculations, recalled from yesterday

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:

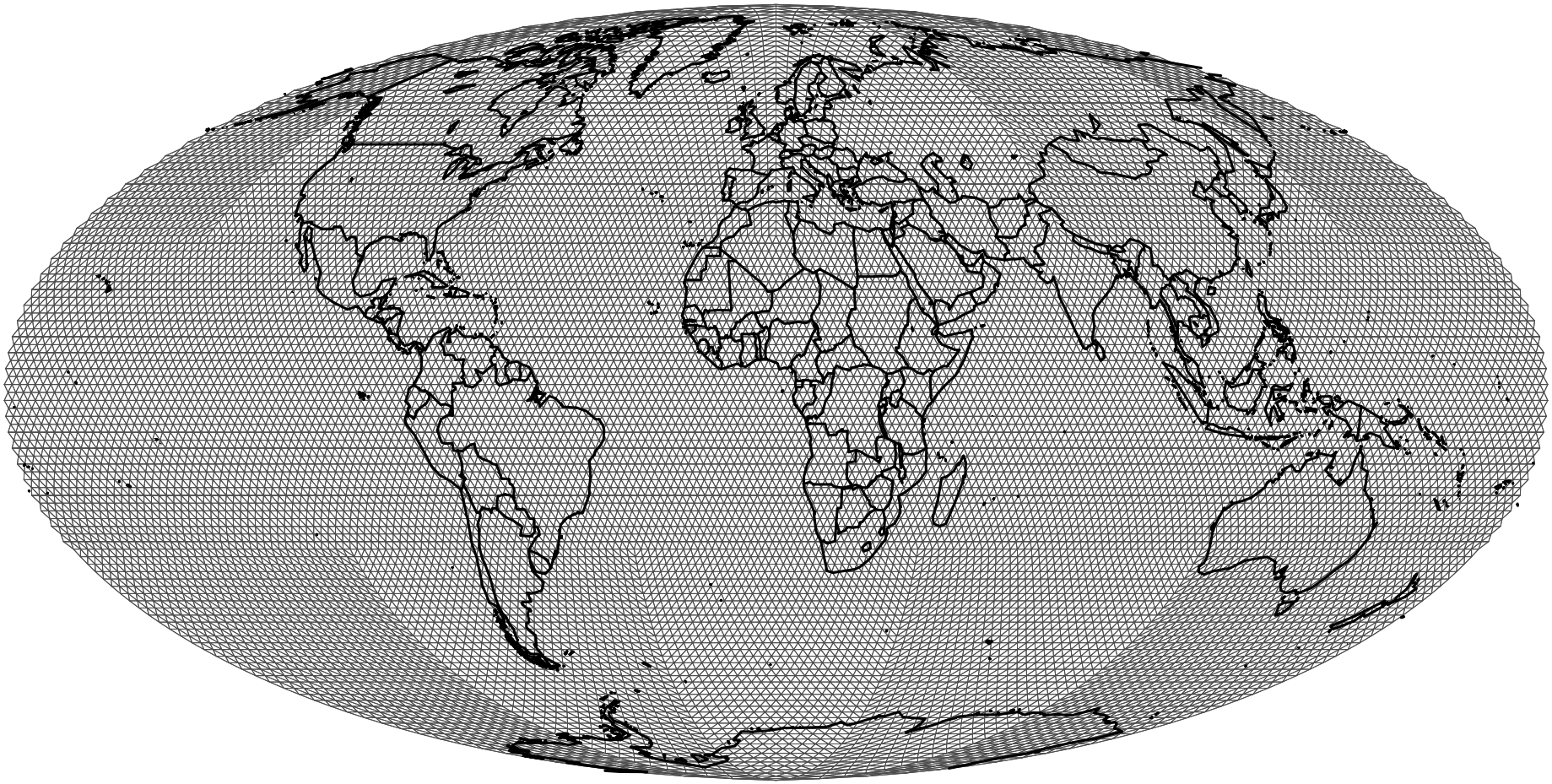
$$\boldsymbol{Q}_{x|y} = \widetilde{\boldsymbol{L}}_{x|y} \widetilde{\boldsymbol{L}}_{x|y}^\top, \qquad \widetilde{\boldsymbol{L}}_{x|y} = \begin{bmatrix} \boldsymbol{L}_t \otimes \boldsymbol{L}_a & \boldsymbol{0} & \boldsymbol{A}^\top \boldsymbol{L}_\epsilon \\ -\boldsymbol{B}^\top \boldsymbol{L}_t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

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

$$\begin{aligned} \boldsymbol{Q}_{x|y}(\boldsymbol{\mu}_{x|y} - \boldsymbol{\mu}_x) &= \widetilde{\boldsymbol{A}}^\top \boldsymbol{Q}_\epsilon (\boldsymbol{y} - \widetilde{\boldsymbol{A}} \boldsymbol{\mu}_x), \\ \boldsymbol{Q}_{x|y}(\boldsymbol{x} - \boldsymbol{\mu}_{x|y}) &= \widetilde{\boldsymbol{L}}_{x|y} \boldsymbol{w}, \quad \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad \text{or} \\ \boldsymbol{Q}_{x|y}(\boldsymbol{x} - \boldsymbol{\mu}_x) &= \widetilde{\boldsymbol{A}}^\top \boldsymbol{Q}_\epsilon (\boldsymbol{y} - \widetilde{\boldsymbol{A}} \boldsymbol{\mu}_x) + \widetilde{\boldsymbol{L}}_{x|y} \boldsymbol{w}, \quad \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \\ \mathrm{Var}(x_i|\boldsymbol{y}) &= \texttt{diag(inla.qinv(}\boldsymbol{Q}_{x|y}\texttt{))} \quad \text{(requires Cholesky)} \end{aligned}$$

The preconditioners for iterative solvers from yesterday are not fully satisfactory.

Finite element mesh



Smaller subregions are partially self-contained sub-problems.

Domain decomposition

- Divide the domain into a collection of overlapping subdomain blocks
- Solve a local problem, e.g. the conditional solution, maintaining coherence by enforcing constraints on overlapping nodes.

Monte Carlo variance reduction for posterior variances

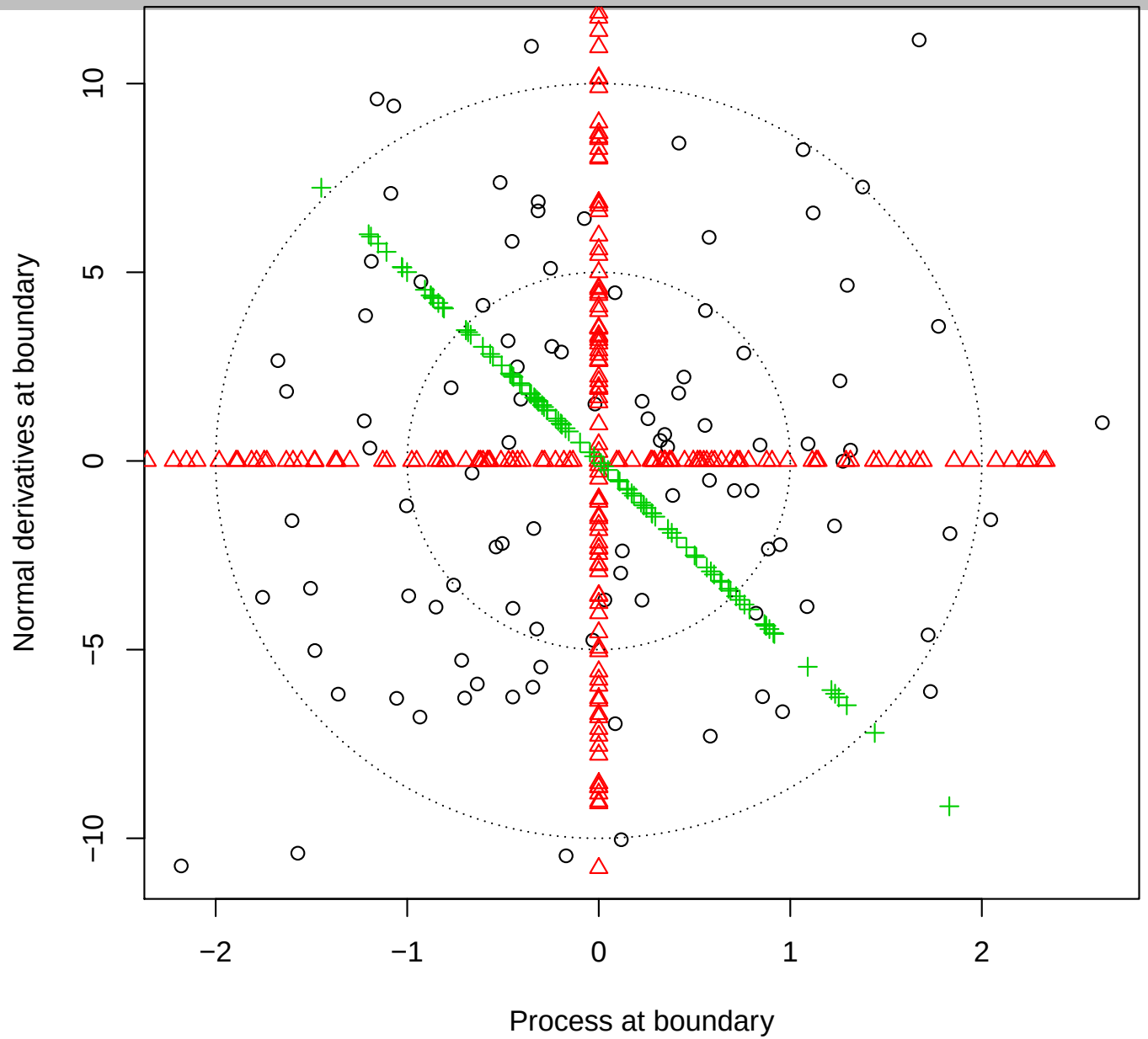
$$\mathsf{E}(\boldsymbol{x}_i \mid \boldsymbol{y}) = \mathsf{E}\left(\mathsf{E}(\boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{x}_{\notin \text{subblock}})\right)$$
$$\mathsf{Var}(\boldsymbol{x}_i \mid \boldsymbol{y}) = \mathsf{Var}\left(\boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{x}_{\notin \text{subblock}}\right) + \mathsf{Var}\left(\mathsf{E}(\boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{x}_{\notin \text{subblock}})\right)$$

Also works for linear combinations, with some complications

Subdomain boundary adjustment (new idea)

- Apply *stochastic boundary correction* for each subdomain
- Solve the full local problem, reusing the appropriate randomness for overlapping subdomains
- Blend the results for overlapping domains.
- Apply this as a preconditioner in an iterative solver

All deterministic boundary conditions are ‘inappropriate’



Stationary stochastic boundary adjustment

Recall the Matérn generating SPDE

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

RKHS inner product on a domain Ω for integer α:

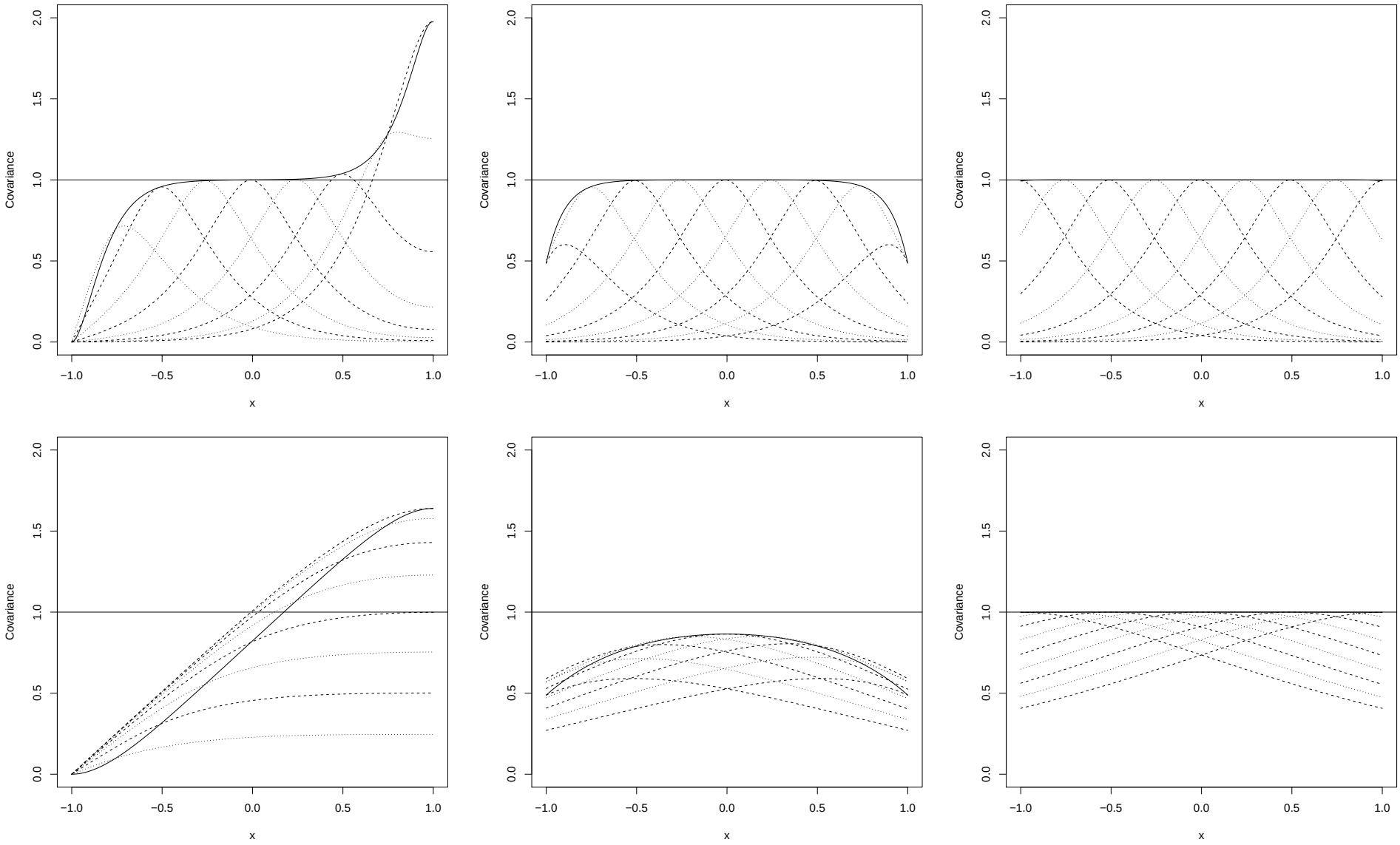
$$\langle f, g \rangle_{H(\Omega)} = \sum_{k=0}^{\alpha} \binom{\alpha}{k} \kappa^{2\alpha-2k} \langle \nabla^k f, \nabla^k g \rangle_{\Omega}$$

Boundary adjusted precision operator on a compact subdomain, where $\mathcal{P}$ projects onto the operator null-space:

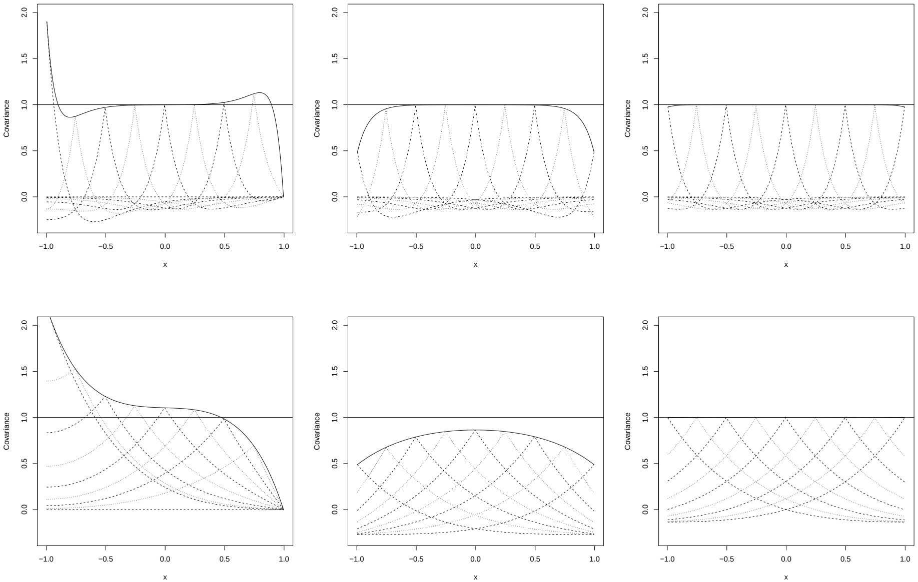
$$\begin{aligned} \mathcal{Q}_{\Omega}(f, g) &= \langle f, g \rangle_{H(\Omega)} - \langle \mathcal{P}f, \mathcal{P}g \rangle_{H(\Omega)} + \mathcal{Q}_{\mathcal{P};\partial\Omega}(\mathcal{P}f, \mathcal{P}g) \\ &= \langle f - \mathcal{P}f, g - \mathcal{P}g \rangle_{H(\Omega)} + \mathcal{Q}_{\mathcal{P};\partial\Omega}(\mathcal{P}f, \mathcal{P}g) \end{aligned}$$

Note that $\mathcal{Q}_{\mathcal{P};\partial\Omega}(\mathcal{P}f, \mathcal{P}g)$ may involve normal derivatives at the boundary.

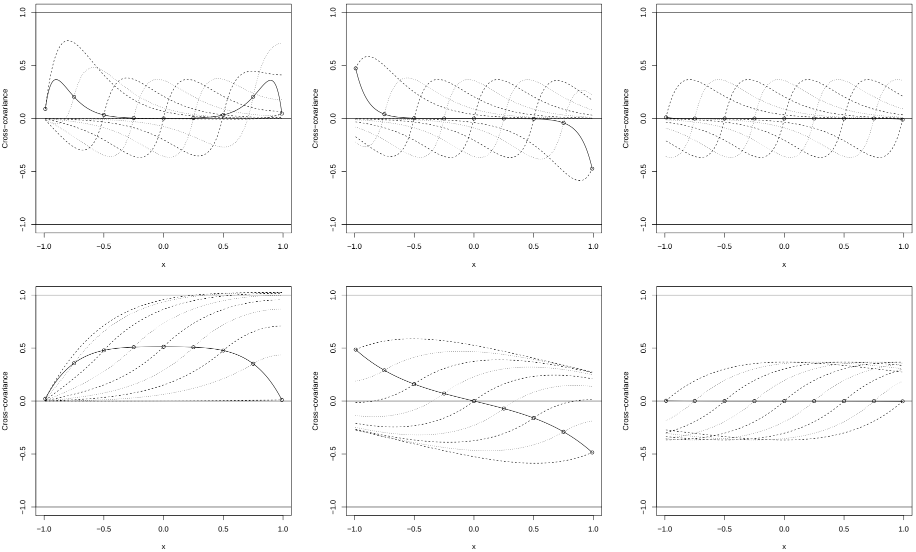
Covariances (D&N, Robin, Stoch) for $\kappa = 5$ and 1



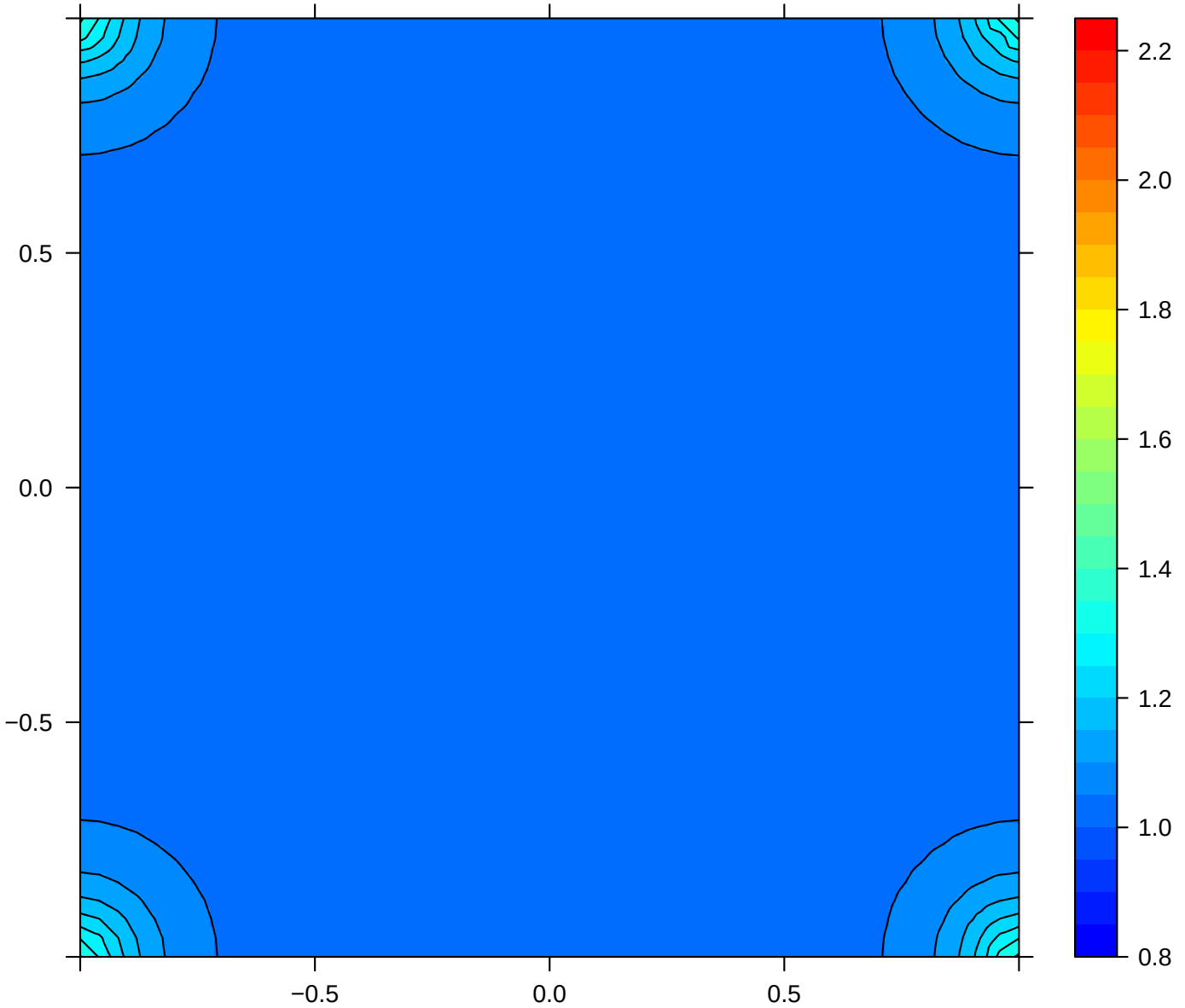
Derivative covariances (D&N, Robin, Stoch) for $\kappa = 5$ and 1



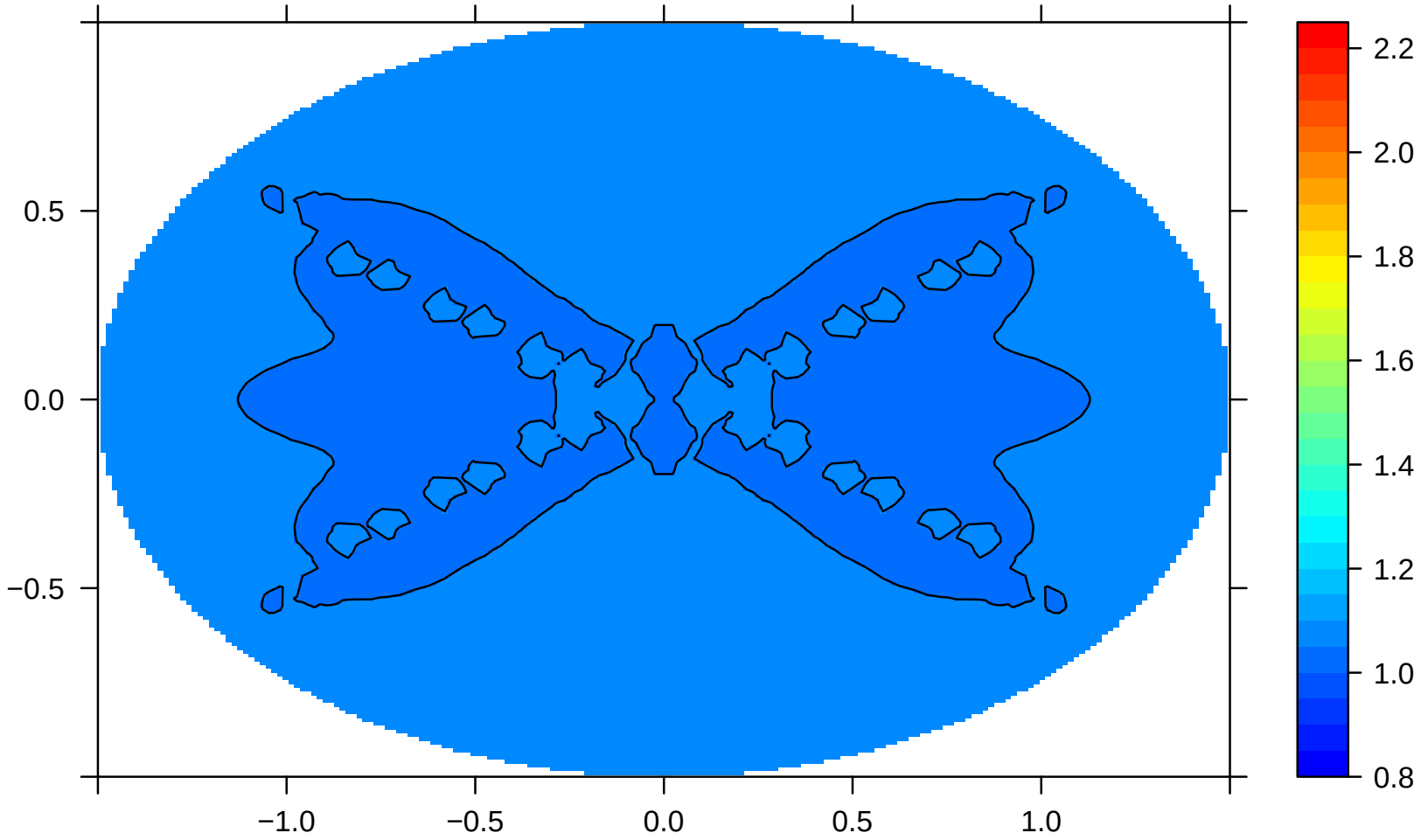
Process-derivative cross-covariances (D&N, Robin, Stoch)



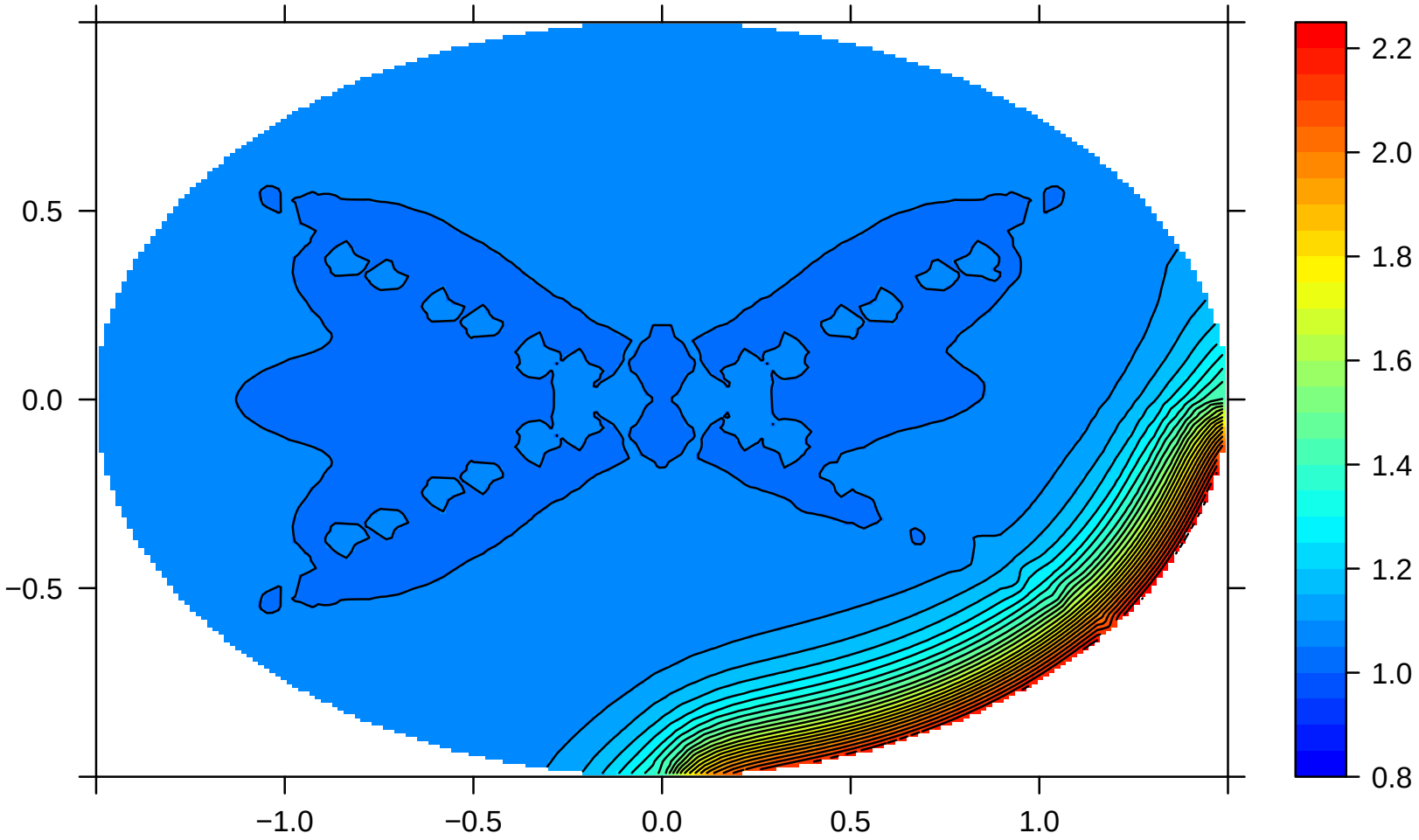
Square domain, stochastic boundary (variances)



Elliptical domain, stochastic boundary (variances)



Elliptical domain, mixed boundary (variances)



References

- Rue, H. and Held, L.: Gaussian Markov Random Fields; Theory and Applications; *Chapman & Hall/CRC*, 2005
- Lindgren, F.: Computation fundamentals of discrete GMRF representations of continuous domain spatial models; preliminary book chapter manuscript, 2014 (covers direct numerical methods)
  `http://people.bath.ac.uk/fl353/tmp/gmrf.pdf`
- Lindgren, F., Rue, H., and Lindström, J.: An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion); *JRSS Series B*, 2011
  Non-CRAN package: `R-INLA` at `http://r-inla.org/`
- Bolin, D. and Lindgren, F.: Excursion and contour uncertainty regions for latent Gaussian models; *JRSS Series B*, 2014, in press. Accepted version at arXiv:1211.3946 and on journal web page.
  CRAN package: `excursions`
  Development: `http://bitbucket.org/davidbolin/excursions`