Science & Technology Facilities Council
University of Leicester
Met Office Hadley Centre
University of Reading
University of Bath
Universität Bern Oeschger Centre Climate Change Research
The University of Edinburgh
DMI
Royal Netherlands Meteorological Institute
Ministry of Infrastructure and the Environment

Building blocks for a statistically advanced daily temperature reconstruction system

Finn Lindgren (`finn.lindgren@ed.ac.uk`)

with Colin Morice, John Kennedy, Christopher Merchant, and the EUSTACE team

THE UNIVERSITY *of* EDINBURGH

IMSC2019, Toulouse, 2019-06-25

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

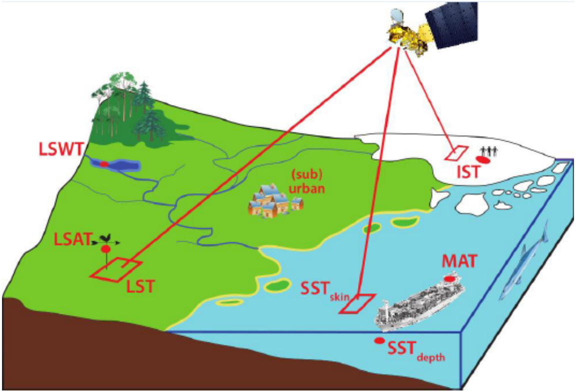
EUSTACE

EUSTACE

EU Surface Temperatures for All Corners of Earth

EUSTACE goal:

Daily estimates of surface air temperature since 1850 across the globe by combining surface and satellite data using novel statistical techniques.



From Merchant et al, Geosci. Instrum. Method. Data Syst., 2, 305-321, 2013

Statistical model and method building blocks

Basic system components

- Multiple *observation sources*, with complex error *uncertainty structure*
- 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
- 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:

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

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

Discrete domain Gaussian Markov random fields (GMRFs)

$\boldsymbol{x} = (x_1, \ldots, x_n) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$ 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).

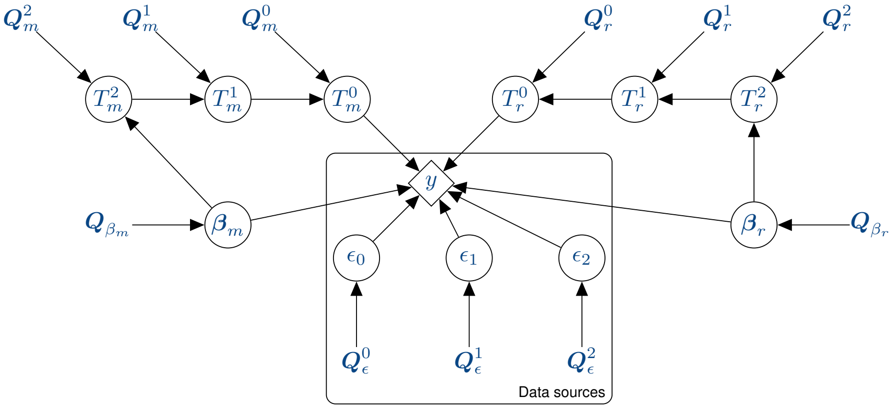
A finite element approximation has structure

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

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.

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

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

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,

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

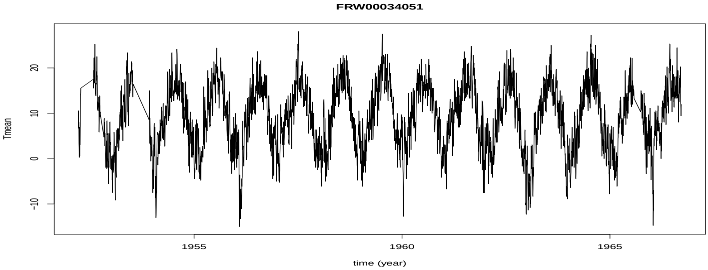
$$y_x^{k,i} = T_m(\mathbf{s}_k, t_i) + \frac{\exp[\widetilde{H}_r^k(t_i)]}{2} T_r(\mathbf{s}_k, t_i) + \epsilon_x^{k,i},$$

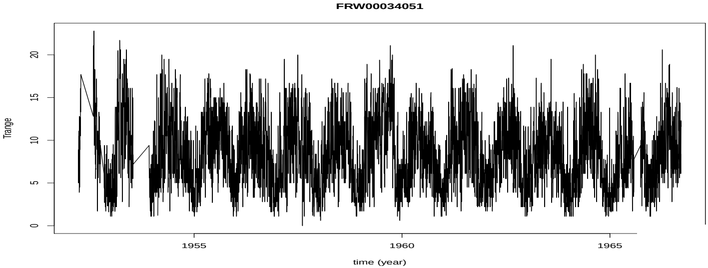
$$y_n^{k,i} = T_m(\mathbf{s}_k, t_i) - \frac{\exp[\widetilde{H}_r^k(t_i)]}{2} T_r(\mathbf{s}_k, t_i) + \epsilon_n^{k,i},$$

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

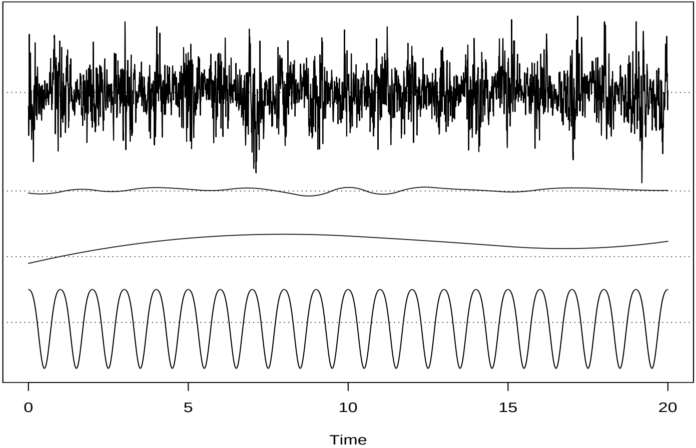
Observed data

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



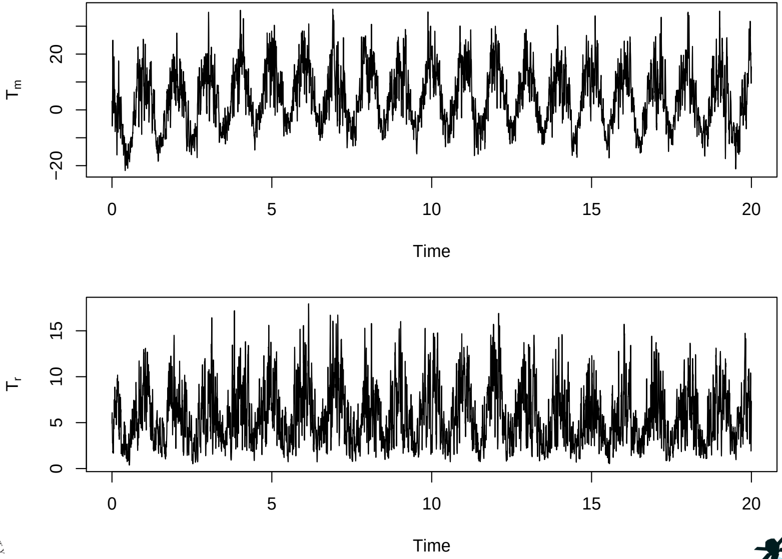


Multiscale model component samples



Combined model samples for T_m and T_r

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



Modelling non-Gaussian quantities

Power tail quantile (POQ) model

The quantile function $F_{\boldsymbol{\theta}}^{-1}(p), p \in [0,1]$, is defined through a quantile blend of left- and right-tailed generalised Pareto distributions:

$$f_\theta^-(p) = \begin{cases} \frac{1-(2p)^{-\theta}}{2\theta}, & \theta \neq 0, \\ \frac{1}{2}\log(2p), & \theta = 0, \end{cases}$$

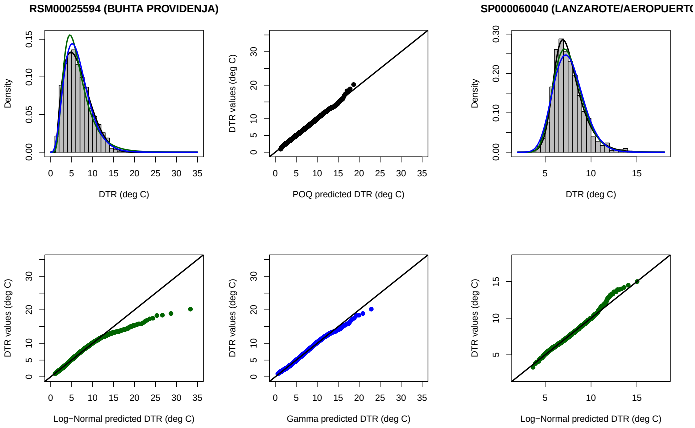
$$f_\theta^+(p) = -f_\theta^-(1-p) = \begin{cases} \frac{(2(1-p))^{-\theta}-1}{2\theta}, & \theta \neq 0, \\ -\frac{1}{2}\log(2(1-p)), & \theta = 0. \end{cases}$$

$$F_{\boldsymbol{\theta}}^{-1}(p) = \theta_0 + \frac{\tau}{2}\left[(1-\gamma)f_{\theta_3}^-(p) + (1+\gamma)f_{\theta_4}^+(p)\right].$$

The parameters $\boldsymbol{\theta} = (\theta_0, \theta_1 = \log\tau, \theta_2 = \operatorname{logit}[(\gamma+1)/2], \theta_3, \theta_4)$ control the median, spread/scale, skewness, and the left and right tail shape.
This model is also known as the *five parameter lambda model* (Gilchrist, 2000).

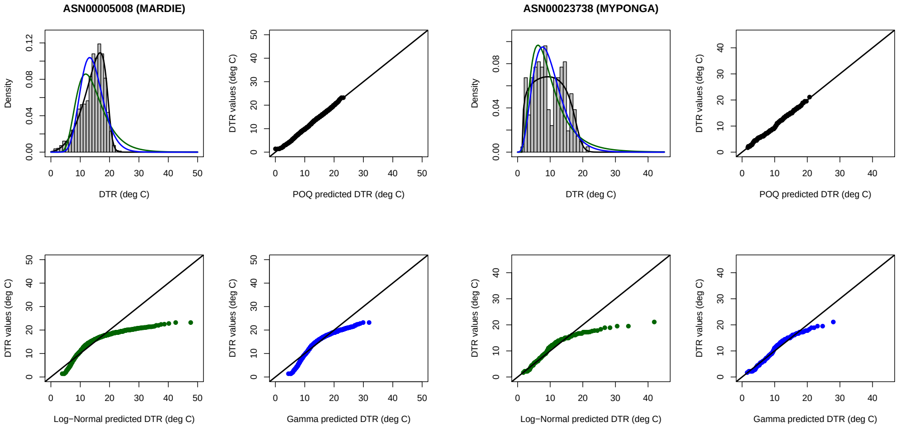
Copula transformation: $G^{-1}[u(\mathbf{s},t)] = F_{\boldsymbol{\theta}(\mathbf{s},t)}^{-1}\{\Phi[u(\mathbf{s},t)]\}$

Diurnal range distributions



For these stations, POQ does a slightly better job than a Gamma distribution.

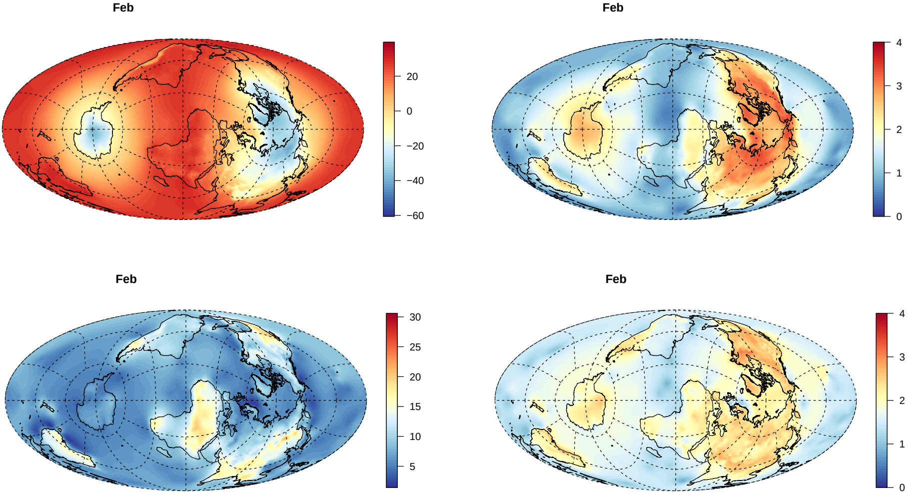
Diurnal range distributions



For these stations only POQ comes close to representing the distributions.

Note: Some of the mixture-like distribution shapes may be an effect of unmodeled station inhomogeneities.

Estimates of median & scale for T_m and T_r



February climatology

(Preliminary estimates, using only in-situ land station data)

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 $\boldsymbol{Q}_x$:

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

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

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

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:

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

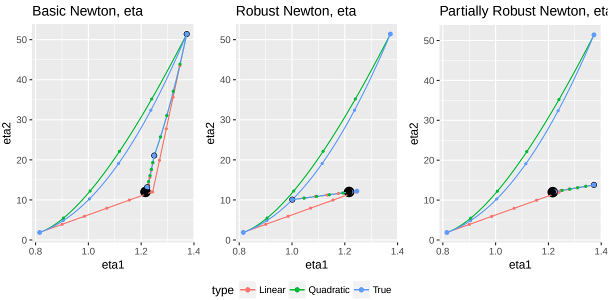
$$\widetilde{\boldsymbol{Q}} = \boldsymbol{Q}_x + \boldsymbol{J}^\top \boldsymbol{Q}_{y|x} \boldsymbol{J}$$

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

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.

Summary

Not covered in this talk:

- Pure conditional block updates risk getting stuck; need for convergence acceleration
- Overlapping space-time blocks for preconditioning
- Non-stationary random field parameter estimation
- Direct&iterative variance calculations to eliminate or reduce Monte Carlo error in the reconstruction uncertainties
- Fast approximate handling of correlated error components

Summary:

- Challenging statistical problem, in both size and complexity
- Approximate calculation techniques allows some of the complexity to be handled with reasonable computational resources
- Close collaboration between climate scientistis, statisticians, and software engineers is essential

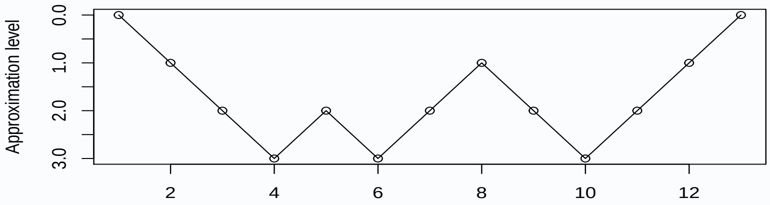
Overlapping blocks and multigrid

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 an additive Schwartz preconditioner is

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

Multigrid and/or approximate multiscale Schur complements



Complications: Schur complements vs conditional block updating

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.

Basic Rao-Blackwellisation of sample estimators

Let $\boldsymbol{x}^{(j)}$ be samples from a Gaussian posterior and let $\boldsymbol{a}^\top \boldsymbol{x}$ be a linear combination of interest. Then, 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 $\boldsymbol{S}$ for each subdomain.

Converting Gaussian to POQ

A POQ copula model

A spatio-temporally dependent Gaussian field $u(\mathbf{s}, t)$ with expectation 0 and variance 1 can be transformed into a POQ field by

$$\widetilde{u}(\mathbf{s}, t) = G^{-1}[u(\mathbf{s}, t)] = F^{-1}_{\boldsymbol{\theta}(\mathbf{s}, t)}[\Phi(u(\mathbf{s}, t)],$$

where the parameters can vary with space and time.

Due to the large size of the problem, we estimate parameters in a two-step procedure:

1. Estimate seasonal POQ and temporal covariance parameters for separate time series
2. With a basic spatial-seasonal random field prior, find the posterior mean parameter field