Statistical Climate Reconstruction Modelling in the EUSTACE Project

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

The University of Edinburgh, Scotland

**with Colin Morice, John Kennedy, and the EUSTACE team,
David Bolin, Haavard Rue, Daniel Simpson, Elias Krainski**

Modern Statistical and Machine Learning Approaches for High-Dimensional Compound Spatial Extremes

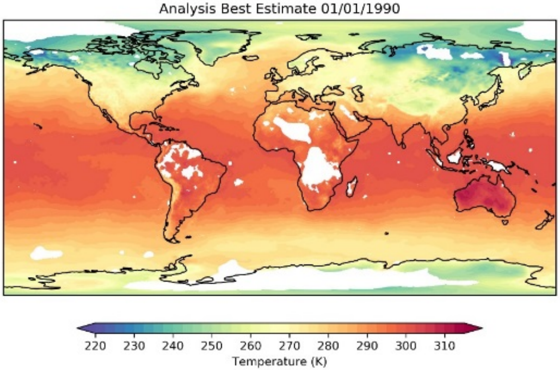
BIRS-IMAG, Granada, 7–12 May 2023

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

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.

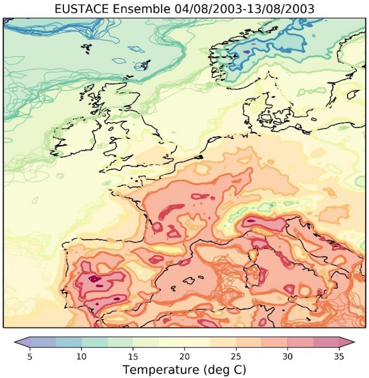


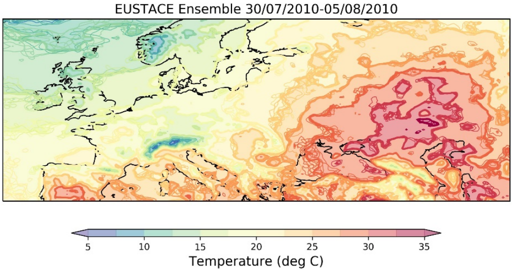
Met Office
Hadley Centre

EUSTACE

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.

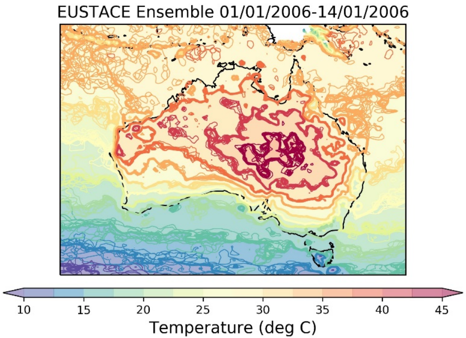


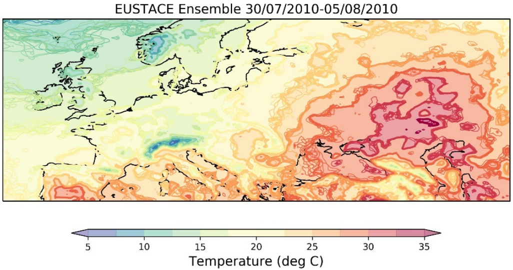


Met Office
Hadley Centre

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

MULTI-SCALE ANALYSIS MODEL

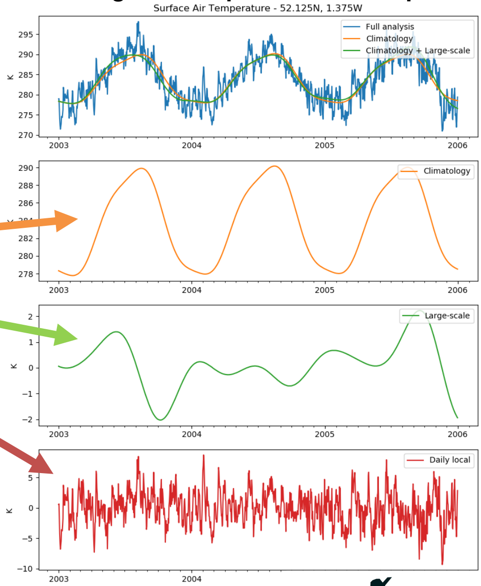
Statistical model for temperature variations and different scales (space and time):

- **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.

Central England Temperature Decomposition



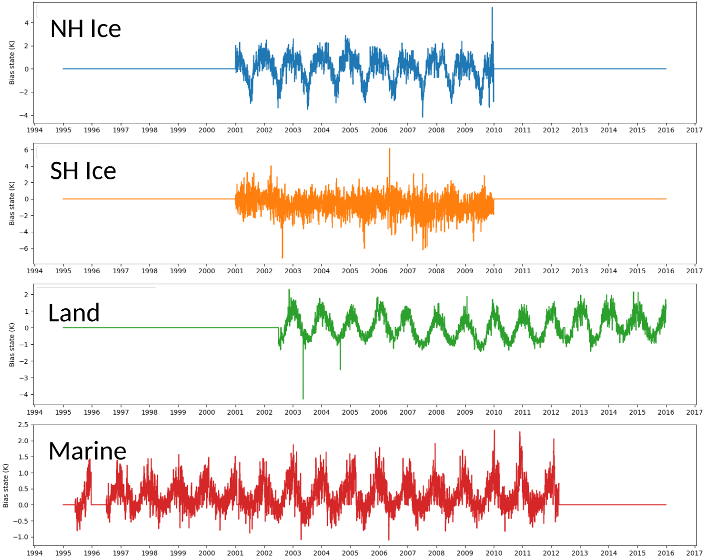
Met Office
Hadley Centre

EUSTACE

SATELLITE BIAS MODELS

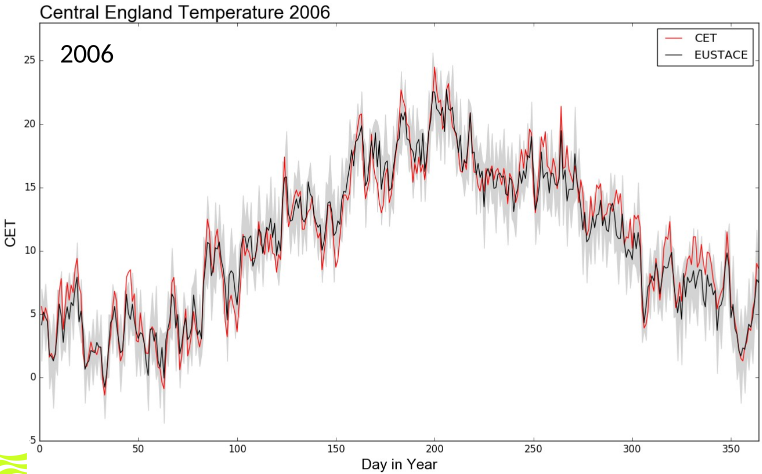
- Simplified model of known error structures in satellite air temperature retrievals:
  - Global/hemispheric systematic bias covariates.
  - Daily estimates of spatially varying bias as a spatial random field.

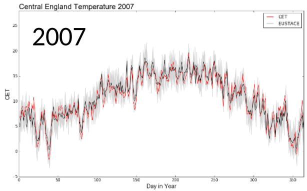
- Estimated jointly with daily temperature variability.

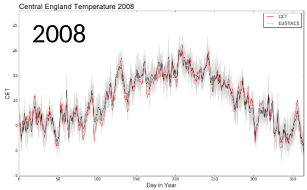


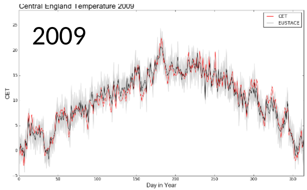
Met Office
Hadley Centre

COMPARING EUSTACE WITH CENTRAL ENGLAND TEMPERATURE









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. 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

GAMs and general kriging

- Linear GAMs with GPs on space and covariates:

$$\eta_i = \sum_k f_k(z_{ik}) + u(\mathbf{s}_i),$$

each $f_k(\cdot)$ and $u(\cdot)$ represented with basis expansions with jointly Gaussian coefficients $\boldsymbol{x}$.

- Linear observations with additive Gaussian observation noise: $\boldsymbol{y} = \boldsymbol{\eta} + \boldsymbol{\epsilon} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}$
- Covariance kriging

$$\boldsymbol{\Sigma}_{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^\top + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$$

$$\mathsf{E}(\boldsymbol{x}|\boldsymbol{y}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^\top\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu})$$

- Precision kriging

$$\boldsymbol{Q}_{\boldsymbol{x}|\boldsymbol{y}} = \boldsymbol{Q}_{\boldsymbol{x}} + \boldsymbol{A}^\top\boldsymbol{Q}_{\boldsymbol{\epsilon}}\boldsymbol{A}$$

$$\mathsf{E}(\boldsymbol{x}|\boldsymbol{y}) = \boldsymbol{\mu} + \boldsymbol{Q}_{\boldsymbol{x}|\boldsymbol{y}}^{-1}\boldsymbol{A}^\top\boldsymbol{Q}_{\boldsymbol{\epsilon}}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu})$$

- Non-Gaussian observations with link function: $\mathsf{E}(y_i|\boldsymbol{\theta}, \boldsymbol{x}) = h(\eta_i)$

Observation level covariance vs latent level precision

- Covariance kriging: linear solve with a $\mathbf{\Sigma}$, $\Sigma_{ij} = \text{Cov}(y_i, y_j)$
  Vecchia approximation:
  $\mathbf{\Sigma}^{-1} \approx \mathbf{L}\mathbf{L}^\top$ for a given observation ordering, and sparse lower triangular L with given sparsity pattern;
  $p(\mathbf{y}|\boldsymbol{\theta}) \approx p(y_1)\prod_{i=2}^{n} p(y_i|\mathbf{y}_{G_i})$, $G_i \subseteq \{1, \ldots, i-1\}$,

$$\mathbf{a}^\top\mathbf{\Sigma}^{-1}\mathbf{a} \approx \sum_i \Big(\sum_{j\in G_i} a_i L_{ij}\Big)^2$$

  $\mathbf{L}$ obtained sequentially from $\mathbf{\Sigma}$ for each observation.
- Precision kriging: linear solve with a $\mathbf{Q}$, $Q_{ij} = \text{Prec}(x_i, x_j|\mathbf{y})$
  $\mathbf{Q} = \mathbf{L}\mathbf{L}^\top$ for a given latent variable ordering, and sparse lower triangular L with the sparsity from $\mathbf{Q}$ plus Cholesky infill.
  The prior $\mathbf{Q}_x$ for SPDE process components is obtained via a local Finite Element construction, giving the model in a chosen finite function space closest to the full model.

Example model: Matérn driven heat equation on the sphere

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

$$\left[\phi\frac{\partial}{\partial t} + (\kappa^2 - \Delta)^{\alpha_s/2}\right]^{\alpha_t} x(\mathbf{s},t)\,\mathrm{d}t = \mathrm{d}\mathcal{E}_{(\kappa^2-\Delta)^{\alpha_e}}(\mathbf{s},t)/\tau$$

For constant parameters, $x(\mathbf{s},t)$ has spatial Matérn covariance (for each t) in a Matérn-Whittle sense on $\mathbb{S}^2$.

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

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

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

THE UNIVERSITY OF EDINBURGH

EUSTACE

Classic and compact INLA methods ($\sim$ description)

- Laplace approximation at the conditional posterior mode $\boldsymbol{x}^*$, and uncertainty integration:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left.\frac{p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{x})}{p(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})}\right|_{\boldsymbol{x}=\boldsymbol{x}^*} \approx \left.\frac{p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{x})}{p_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})}\right|_{\boldsymbol{x}=\boldsymbol{x}^*} = \widehat{p}(\boldsymbol{\theta}|\boldsymbol{y})$$

$$p(x_i|\boldsymbol{y}) = \int p(x_i|\boldsymbol{\theta},\boldsymbol{y})p(\boldsymbol{\theta}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{\theta} \approx \sum_k \widehat{p}(x_i|\boldsymbol{\theta}^{(k)},\boldsymbol{y})\widehat{p}(\boldsymbol{\theta}^{(k)}|\boldsymbol{y})w_k = \widehat{p}(x_i|\boldsymbol{y})$$

- Let $\widehat{\boldsymbol{\mu}} = \mathsf{E}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ and $\boldsymbol{Q}_\epsilon = -\nabla_x\nabla_x^\top \log p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{x}^*)$
- Classic method: Laplace approximation of each $\widehat{p}(x_i|\boldsymbol{\theta},\boldsymbol{y})$, and
$$\left\{\begin{bmatrix}\boldsymbol{A}\boldsymbol{x}\\ \boldsymbol{x}\end{bmatrix}\middle|\boldsymbol{\theta},\boldsymbol{y}\right\} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{A}\widehat{\boldsymbol{\mu}}\\ \widehat{\boldsymbol{\mu}}\end{bmatrix}, \begin{bmatrix}\boldsymbol{Q}_\epsilon+\delta\boldsymbol{I} & -\delta\boldsymbol{A}\\ -\delta\boldsymbol{A}^\top & \boldsymbol{Q}_x+\delta\boldsymbol{A}^\top\boldsymbol{A}\end{bmatrix}^{-1}\right)$$, with $\delta \gg 0$
- Compact method: Variational approximation of $\widehat{p}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$, and
$$\{\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}\} \sim \mathcal{N}\left(\widehat{\boldsymbol{\mu}}, [\boldsymbol{Q}_x+\boldsymbol{A}^\top\boldsymbol{Q}_\epsilon\boldsymbol{A}]^{-1}\right)$$

Before satellites you had to go measure in person

"The Discovery", Dundee, Scotland (Photos: Finn Lindgren, August 2022)

Hydrology lab from the 1925-27 Antarctic ocean expedition

"The Discovery", Dundee, Scotland (Photos: Finn Lindgren, August 2022)

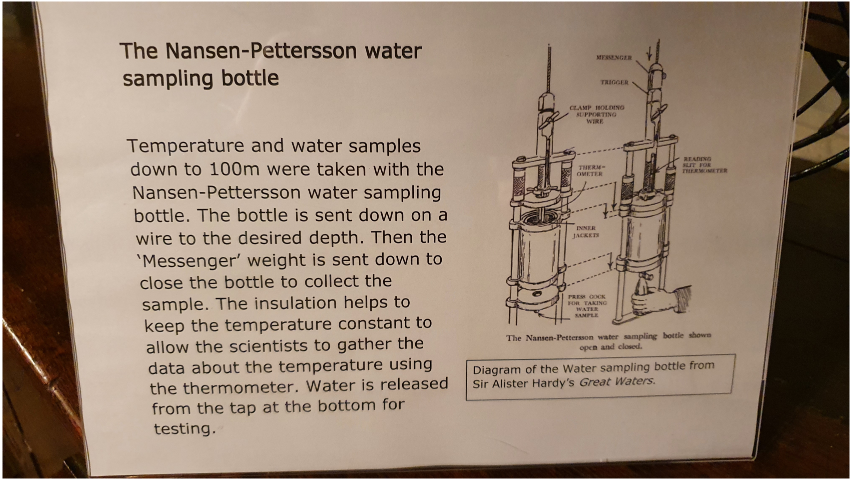
THE UNIVERSITY OF EDINBURGH

EUSTACE

What’s that in the corner?

"The Discovery", Dundee, Scotland (Photos: Finn Lindgren, August 2022)

It’s a Nansen-Pettersson water sampling bottle!



"The Discovery", Dundee, Scotland (Photos: Finn Lindgren, August 2022)

Station observation & homogenisation model

Daily mean air temperature measurements

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

$$y_x^{k,i} = T_m(\mathbf{s}_k, t_i) + \widetilde{H}_{r,m}^k(t_i) + \frac{\widetilde{H}_{r,r}^k(t_i)}{2} T_r(\mathbf{s}_k, t_i) + \epsilon_x^{k,i},$$

$$y_n^{k,i} = T_m(\mathbf{s}_k, t_i) + \widetilde{H}_{r,m}^k(t_i) - \frac{\widetilde{H}_{r,r}^k(t_i)}{2} T_r(\mathbf{s}_k, t_i) + \epsilon_n^{k,i},$$

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

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:

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

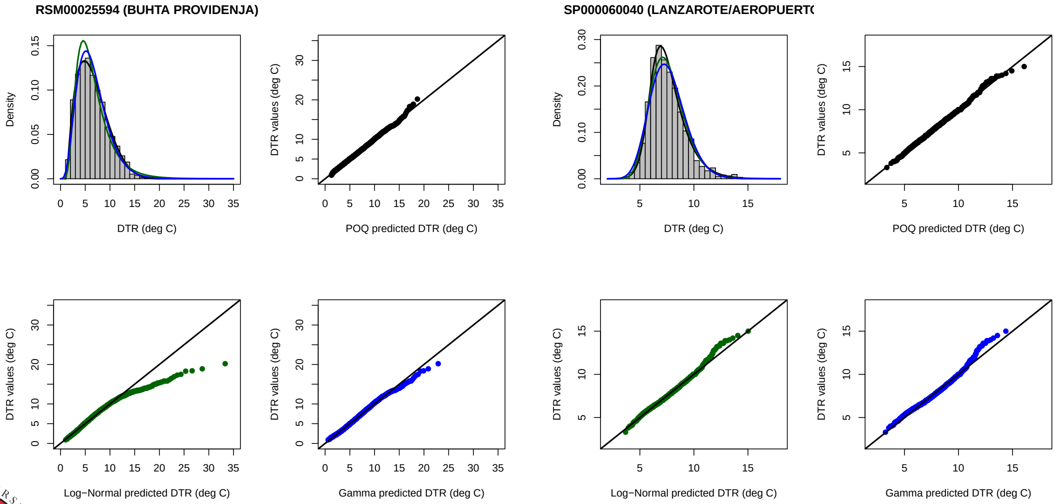
$$f_\theta^+(p) = -f_\theta^-(1-p) = \begin{cases} \frac{(2(1-p))^{-\theta}-1}{2\theta}, & \theta \neq 0, \\ -\frac{1}{2}\log(2(1-p)), & \theta = 0. \end{cases}$$

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

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).

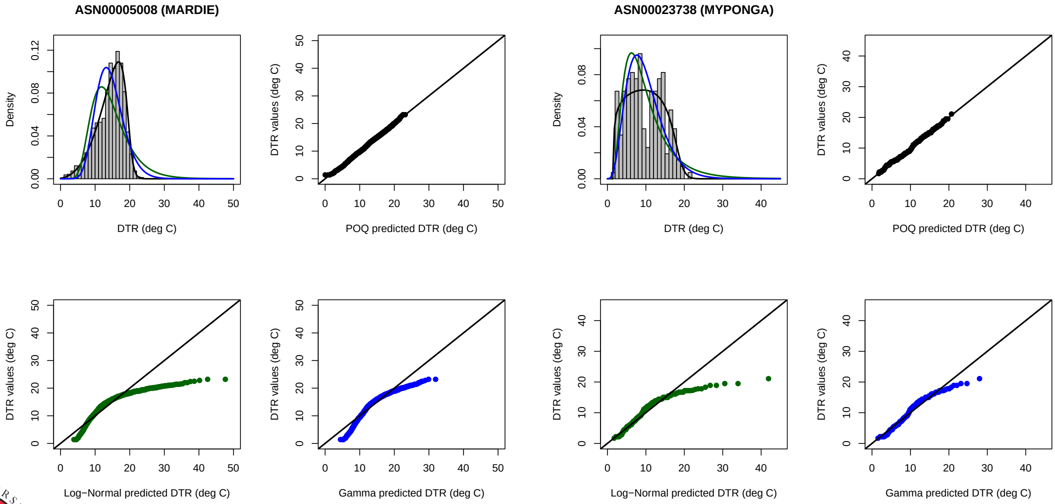
Transformation: $T_r(\mathbf{s},t) = F_{\boldsymbol{\theta}(\mathbf{s},t)}^{-1}\{\Phi[u(\mathbf{s},t)]\}$, $\mathsf{E}[u(\mathbf{s},t)] = 0$, $\mathsf{Var}[u(\mathbf{s},t)] = 1$

Diurnal range distributions



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

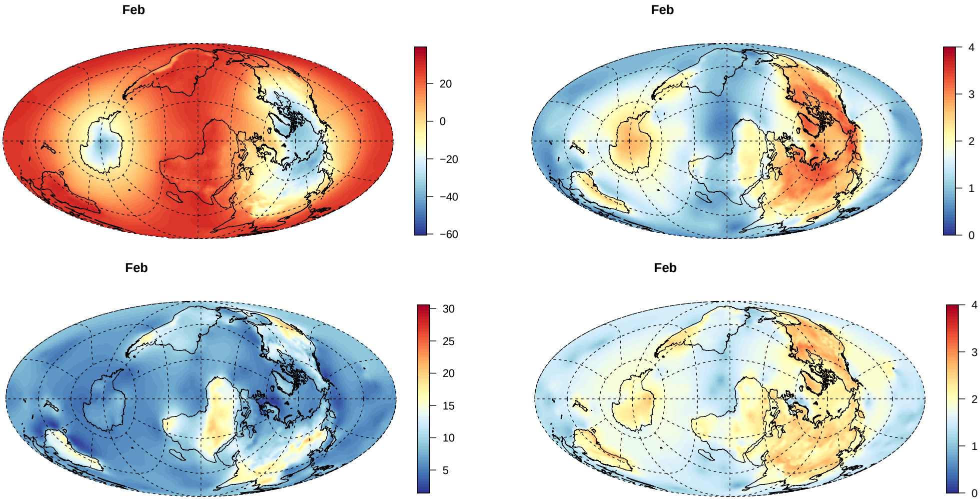
Diurnal range distributions



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

Note: Some shapes may be due to unmodeled station inhomogeneities.

Estimates of median & scale for T_m and T_r



February climatology

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

EUSTACE

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 $\boldsymbol{Q}_x$:

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

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

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

Non-linear and/or non-Gaussian observations

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

$$(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}) \overset{\text{approx}}{\sim} \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{Q}}^{-1}) \qquad \text{(INLA posterior from } \overline{h}(\boldsymbol{x}) = h(\boldsymbol{\mu}^*) + \boldsymbol{J}(\boldsymbol{x} - \boldsymbol{\mu}^*)\text{)}$$

$$\widetilde{\boldsymbol{Q}} = \boldsymbol{Q}_x + \boldsymbol{J}^\top \boldsymbol{Q}_{y|x} \boldsymbol{J} \qquad \text{(Generally: } \boldsymbol{Q}_x - \nabla_x \nabla_x^\top \log p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\text{)}$$

$$\boldsymbol{\mu}^*_{\text{new}} = \boldsymbol{\mu}^* + (\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^*) \cdot \operatorname*{argmin}_{a>0} \|\overline{h}(\widetilde{\boldsymbol{\mu}}) - h(\boldsymbol{\mu}^* + (\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^*)a)\|$$

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, 2015, https://www.maths.ed.ac.uk/~flindgre/cuso2019/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
  R packages: INLA (http://r-inla.org/)
  and inlabru (https://inlabru-org.github.io/inlabru/)
- Lindgren, F., Bolin, D., and Rue, H.: The SPDE Approach for Gaussian and Non-Gaussian Fields: 10 Years and Still Running; *Spatial Statistics, Special Issue: The Impact of Spatial Statistics*, 50:100599, 2022. https://arxiv.org/abs/2111.01084
- Lindgren, F., Haakon Bakka, David Bolin, Elias Krainski, Håvard Rue: A diffusion-based spatio-temporal extension of Gaussian Matérn fields, arXiv 2020–2023. https://arxiv.org/abs/2006.04917
- Video illustrating the results, produced by Philip Brohan:
  https://twitter.com/philipbrohan/status/1253411283598073867
  https://player.vimeo.com/video/403663259
- Links to EUSTACE project reports and data:
  https://www.eustaceproject.org/

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.

THE UNIVERSITY OF EDINBURGH

EUSTACE