Hierarchical Models for Environmental Extremes

Lee Fawcett, Dave Walshaw
and Jonathan Atyeo

Newcastle University, Newcastle upon Tyne, U.K.

NBBC11: Turku, Finland, June 2011

Structure of this talk

1. Background and motivating examples

- Example: Extreme wind speeds in the U.K.
- Example: Extreme rainfall in the U.K.
- Statistical modelling of extremes

2. A hierarchical model for extreme wind speeds

- Site and seasonal variation
- Temporal dependence
- Model structure and inference

3. A Hierarchical model for extreme rainfall

- Spatial dependence
- Regional variation and temporal trends
- Model structure and inference

1.1 Background and motivation: extreme wind speeds

In the U.K., the *British Standards Institution* produce contour maps displaying strength requirements for structures based on "once–in–50–year gust speeds".

This is known as the 50–year return level gust.

The maps themselves are the result of simple extreme value analyses carried out on medium to long term records collected at stations in the U.K.

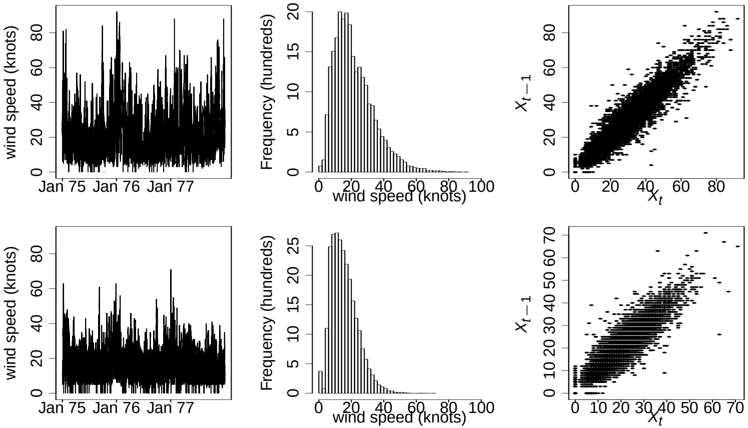
1.1 Background and motivation: extreme wind speeds

During storms in 1987, 2002 and 2005, gust speeds exceeded the 200–year return level.

- Perhaps building codes should be revised?
- Or maybe the estimation procedure is inappropriate...

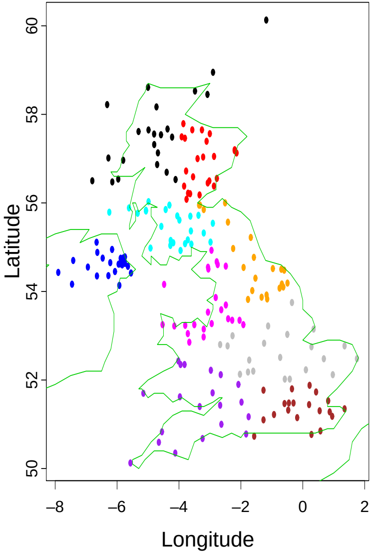
1.1 Background and motivation: extreme wind speeds

Wind speed stations: Central/Northern England



Exploratory analysis of wind speed data

1.2 Background and motivation: extreme rainfall

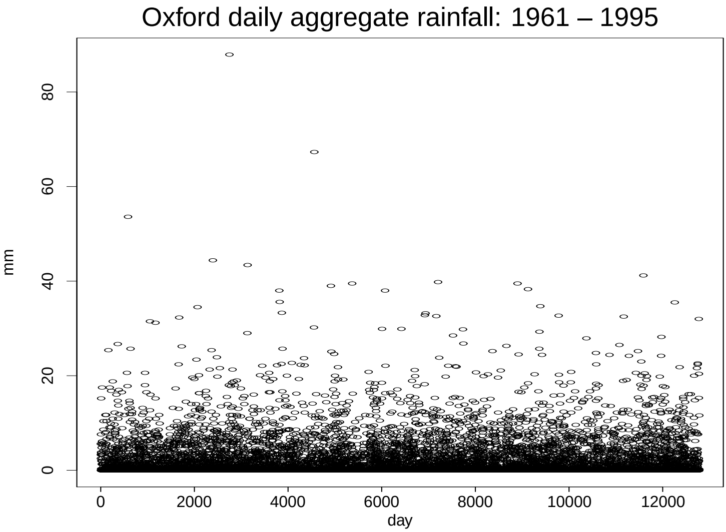


Regions

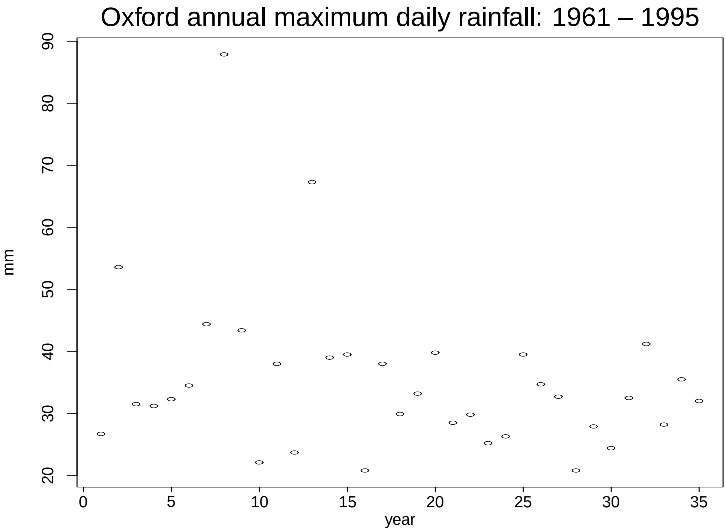
Black:	Northern Scotland (NS)
Red:	Eastern Scotland (ES)
Cyan:	Southern Scotland (SS)
Blue:	Northern Ireland (NI)
Magenta:	North West England (NWE)
Orange:	North East England (NEE)
Grey:	Central East England (CEE)
Purple:	South West England (SWE)
Brown:	South East England (SEE)

Rainfall stations: UK

1.2 Background and motivation: extreme rainfall



1.2 Background and motivation: extreme rainfall



1.3 Background and motivation: modelling extremes

Let $X_1, X_2, \ldots, X_n$ be a stationary sequence of random variables with common distribution function F, and let

$$M_n = \max\{X_1, \ldots, X_n\}.$$

Then, as $n \to \infty$,

$$\Pr(M_n \leq x) \approx F^{n\theta}(x),$$

where $\theta \in (0, 1)$ is the extremal index.

- As $\theta \to 0$: increasing dependence in the extremes
- For $\theta = 1$: we have an independent process.

1.3.1 The "block maxima" approach

Idea: Find approximate families of models for $G = F^n$ as $n \to \infty$.

For independent extremes, i.e. where $\theta = 1$, this leads to the Generalised Extreme Value (GEV) distribution, where

$$G(y;\mu,\sigma,\xi) = \exp\left\{-\left[1+\xi\left(\frac{y-\mu}{\sigma}\right)\right]_{+}^{-1/\xi}\right\},$$

where $a_{+} = \max(0, a)$ and μ, $\sigma > 0$ and ξ are location, scale and shape parameters (respectively).

Allowing for dependence has the effect of powering the limit distribution G by θ – again giving a GEV, albeit with different location and scale parameters.

1.3.2 A threshold–based approach

Leadbetter *et al.* (1983) show that for large u and for $\theta = 1$, the distribution of $(X - u)$, conditional on $X > u$, is approximately Generalised Pareto (GP), with distribution function

$$G(y; \bar{\sigma}, \xi) = 1 - \left(1 + \frac{\xi y}{\bar{\sigma}}\right)_{+}^{-1/\xi},$$

where $\bar{\sigma} = \sigma + \xi(u - \mu)$.

Graphical procedures can be used to select the threshold u; the GP distribution is then used to model excesses over u.

This approach is often preferred over the block maxima approach owing to the inclusion of more extremes.

1.3.3 Modelling issues

Temporal dependence

- Not an issue for the GEV, provided long–range dependence is negligible
- Workaround in the threshold approach: filter out a set of independent extremes and fit the GP to these

Seasonal variability – workaround: use data from seasons giving the largest extremes, and within which we can assume stationarity – ignore all other seasons

Site–by–site variability – workaround: independent site–by–site analyses

The aim of our work is to *exploit* such meteorological structure – not use "workarounds" – to give more informed inferences!

2. Extreme wind speeds in the U.K.

Main aim:

- Provide accurate, precise estimates of return levels, by...
- ...getting as much out of our data as possible!

This means:

- Using a threshold approach to minimise data wastage
  - Avoid filtering – also wasteful
  - Will need to think about temporal dependence
- Using extremes from *all* seasons – perhaps by quantifying seasonal effects
- Pooling extremes from *all* sites into a single analysis to allow information–sharing
- Being Bayesian – enabling us to augment our analysis with prior information, as well as obtain predictive return level estimates

2.1 Site and seasonal variation

We use a site– and seasonally–varying GP distribution for threshold excesses above $u_{m,j}$, yielding

$$(\tilde{\sigma}_{m,j}, \xi_{m,j}), \qquad m, j = 1, \ldots, 12,$$

where m and j are indices of season and site (respectively).

The GP scale parameter $\bar{\sigma}$ is *threshold–dependent*; to allow comparisons across different sites and seasons, we work with

$$\tilde{\sigma}_{m,j} = \bar{\sigma}_{m,j} - \xi_{m,j} u_{m,j}.$$

- The scale parameter is now *threshold–independent*
- In a Bayesian setting, this is desirable: this allows us to specify prior information about $(\tilde{\sigma}_{m,j}, \xi_{m,j})$ without having to worry about threshold–dependency

2.2 Temporal dependence

Previous work in Fawcett and Walshaw (2006) indicates the suitability of a first–order Markov assumption for consecutive extreme wind speeds.

Thus, the likelihood for our series of extremes in season m at site j is given by

$$L(\underline{\psi}; y_1, \ldots, y_n) = \prod_{i=1}^{n-1} f(y_i, y_{i+1}; \underline{\psi}) \Bigg/ \prod_{i=2}^{n-1} f(y_i; \underline{\psi}).$$

- Contributions to the denominator, for the region $(u_{m,j}, \infty)$, are given by the GP density
- For contributions to the numerator we need an appropriate bivariate extreme value model

2.2 Temporal dependence

Fawcett and Walshaw (2006) try several bivariate extreme value models for the wind speed extremes.

- Allowing for *asymmetry* in the dependence structure between (y_i, y_{i+1}) is unnecessary
- Of the *symmetric* models tried, the logistic model gave the best fit:

$$F(y_i, y_{i+1}) = 1 - \left(Z(y_i)^{-1/r} + Z(y_{i+1})^{-1/r}\right)^r, \qquad y_i, y_{i+1} > u_{m,j},$$

  - The transformation Z ensures the margins are of GP form
  - r is a dependence parameter – independence and complete dependence are obtained when $r = 1$ and $r \searrow 0$

2.3 Hierarchical structure

We specify the following hierarchical model for our wind speed data:

$$\begin{aligned} \eta_{m,j} = \log(\tilde{\sigma}_{m,j}) &= \gamma_{\eta}^{(m)} + \kappa_{\eta}^{(j)} \\ \xi_{m,j} &= \gamma_{\xi}^{(m)} + \kappa_{\xi}^{(j)} \qquad \text{and} \\ r_j &= \kappa_r^{(j)}, \end{aligned}$$

where γ and κ represent seasonal and site effects respectively.

Previous studies in Fawcett and Walshaw (2008) show little variation in r from month to month, but variations between sites.

2.3 Hierarchical structure

Notation: ϵ_ω used generically for any of γ_η, γ_ξ, κ_η, or κ_ξ.

Random effects distributions for $\eta_{m,j}$ and $\xi_{m,j}$:

$$\epsilon_\omega \sim N_0(\alpha_\omega, \tau_\omega).$$

A conditional autoregressive model structure is used for all seasonal effects:

$$\alpha_\omega^{(m)} = \frac{1}{2}\left(\alpha_\omega^{(m-1)} + \alpha_\omega^{(m+1)}\right), \qquad m = 1, \ldots 12.$$

For site effects, we use:

$$\begin{aligned} \alpha_\omega^{(j)} &\sim N_0(b_\omega, c_\omega), \qquad \text{and} \\ \kappa_r^{(j)} &\sim U(0,1), \qquad j = 1, \ldots, 12. \end{aligned}$$

To retain conjugacy, we also use:

$$\tau_\omega \sim \mathrm{Ga}(d_\omega, f_\omega).$$

2.4 Inference

Estimation of the random effects model is made via a Metropolis–within–Gibbs MCMC algorithm

- We sample from the full conditionals of the random effects means and precisions α_ω and τ_ω
- The complexity of the GP likelihood requires a Metropolis step for the random effects themselves

The sampler yields approximate draws from the posteriors for

- the 12 site effects parameters for each of $\log(\tilde{\sigma}_{m,j})$, $\xi_{m,j}$ and the logistic dependence parameter r_j;
- the 12 seasonal effects for each of $\log(\tilde{\sigma}_{m,j})$ and $\xi_{m,j}$.

2.5 Results: Random effects and GPD parameters

	Bradfield, January Mean (st. dev.) MLE (s.e.)	Nottingham, July Mean (st. dev.) MLE (s.e.)
$\gamma_\eta^{(m)}$	1.891 (0.042)	1.294 (0.042)
$\gamma_\xi^{(m)}$	0.021 (0.018)	0.002 (0.018)
$\kappa_\eta^{(j)}$	0.367 (0.044)	–0.121 (0.041)
$\kappa_\xi^{(j)}$	–0.105 (0.020)	–0.059 (0.017)
$\kappa_r^{(j)}$	0.385 (0.009)	0.400 (0.011)
$\tilde{\sigma}_{m,j}$	7.267 (0.211) 8.149 (0.633)	3.234 (0.061) 2.914 (0.163)
$\xi_{m,j}$	–0.084 (0.015) –0.102 (0.055)	–0.057 (0.013) 0.018 (0.044)
α_j	0.385 (0.009) 0.368 (0.012)	0.400 (0.011) 0.412 (0.020)

Bayesian random effects analysis

2.5 Results: Return level inference

For each site j the annual exceedance rate of q_r is given by:

$$\sum_{m=1}^{12} \left\{1 - F_{m,j}(q_r)^{h_{m,j}\theta_j}\right\}, \qquad m = 1, \ldots, 12,$$

where

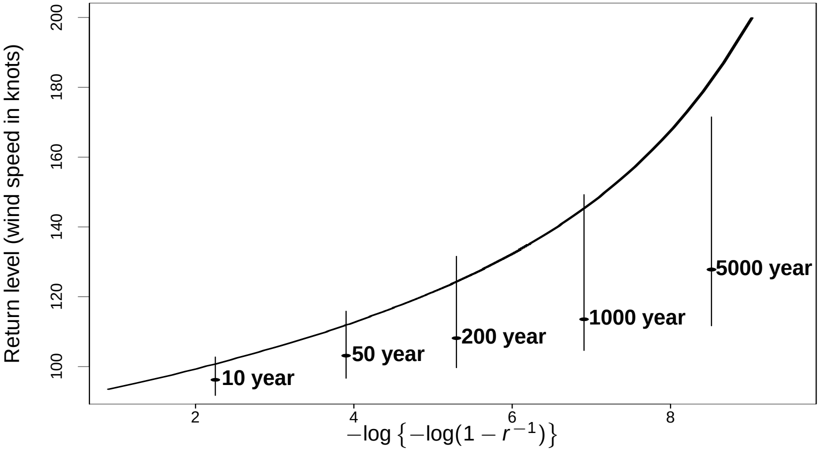
- $\{1 - F_{m,j}(q_r)^{h_{m,j}\theta_j}\}$ is the annual exceedance rate of q_r in month m;
- $F_{m,j}$ is the GPD distribution function in month m with parameters $\tilde{\sigma}_{m,j}$ and $\xi_{m,j}$;
- $h_{m,j}$ is the number of hours in month m, and
- the extremal index θ_j is implicitly defined through the value of the logistic dependence parameter α_j at site j.

2.5 Results: Return level inference

	Return Period (years)			
	10	50	200	1000
Hierarchical model	96.887 (0.982)	103.463 (1.333)	112.518 (2.023)	128.128 (2.691)
Maximum likelihood	96.745 (2.864)	103.236 (5.930)	108.152 (8.786)	113.306 (12.219)
Predictive	104.392	113.089	119.957	127.338

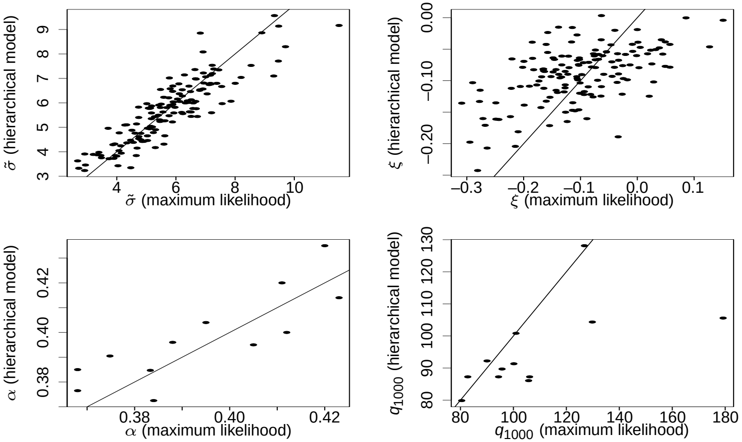
Return levels for Bradfield (knots)

2.5 Results: Return level inference



Predictive return level curve for Bradfield

2.5 Results: “Shrinkage”

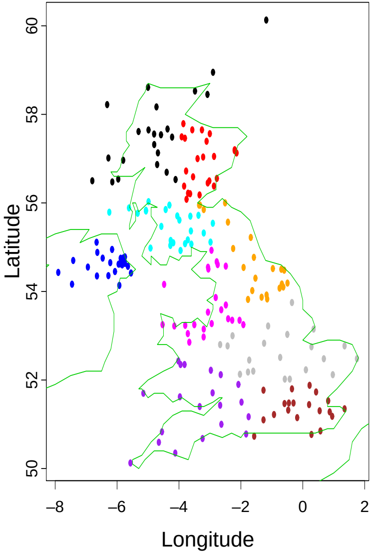


Posterior means against maximum likelihood estimates

3. Extreme rainfall in the U.K.

3. Extreme rainfall in the U.K.

3. Extreme rainfall in the U.K.



Regions

Black:	Northern Scotland (NS)
Red:	Eastern Scotland (ES)
Cyan:	Southern Scotland (SS)
Blue:	Northern Ireland (NI)
Magenta:	North West England (NWE)
Orange:	North East England (NEE)
Grey:	Central East England (CEE)
Purple:	South West England (SWE)
Brown:	South East England (SEE)

Rainfall stations: UK

3.1 Contrasts with the wind speeds problem

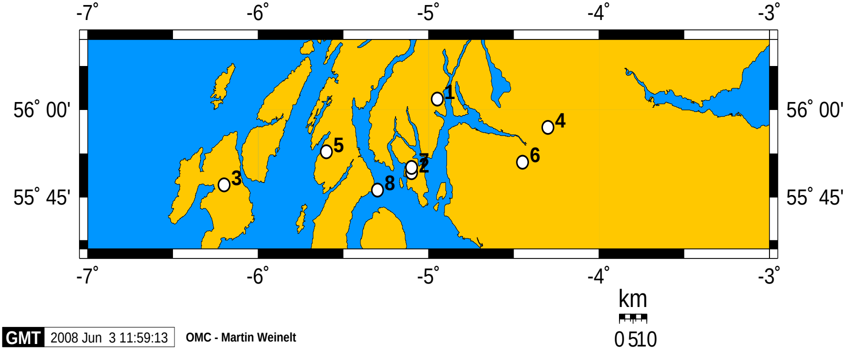
Data are daily aggregate rainfalls for 204 locations over 9 regions, 1961 - 2000. We focus on annual maxima.

Modelling issues:

- multiple geographic regions with differing climates;
- significant dependence across sites within regions;
- annual maxima - no seasonal effects;
- annual maxima - no temporal dependence;
- data range 1961 - 2000: climate change?

Example

Suppose we want to study the joint extremes of daily rainfall accumulations at the network of 8 sites shown below.



3.2 Modelling inter-site dependence within a region

- We have already seen that the GEV is the class of limit distributions for maxima at a single location.
- It is natural to consider a multivariate analogue for the multivariate annual maxima.
- However detailed investigation of the multivariate behaviour of the extreme rainfall suggests asymptotic *independence*.
- *However,* significant dependence persists to the levels of the annual maxima.
- We employ a pragmatic solution, which is a departure from asymptotic theory . . .

3.3 The multivariate Gaussian tail model

We use a multivariate Gaussian tail model, as suggested by Bortot *et al.* (2000):—

We model the annual maxima for each of the p sites within a region using a GEV distribution, and transform these margins to standard Gaussian.

We then model the joint distribution of these transformed maxima using a p–dimensional Normal distribution with standard margins.

The dependence structure now has the property of asymptotic independence, which appears to be appropriate.

3.3 The multivariate Gaussian tail model

Consider a p dimensional random variable $\mathbf{X_i} = (X_{i,1}, ..., X_{i,p})$ and let $M_n = (\mathbf{M}_{n,1}, \ldots, \mathbf{M}_{n,p})$, where $\mathbf{M}_{n,j} = \max_{1 \leq i \leq n} \{X_{i,j}\}$.

Note that M_n is not necessarily one of the original $\mathbf{X}_i$ vectors.

We assume the p-dimensional block maxima random variable M_n to have marginal distribution G_j, $j = 1, \ldots, p$, where the G_j have the GEV distribution.

3.3 The multivariate Gaussian tail model

Denoting by $\Phi(.)$ and $\Phi_p\,(.;\Sigma)$ the standard Normal distribution function and the standard p-dimensional Normal distribution function with correlation matrix Σ respectively, we define the distribution function for M_n as

$$G(x_1,\ldots,x_p) = \Phi_p\left\{\Upsilon_1(x_1),\ldots,\Upsilon_p(x_p);\Sigma\right\}$$

where

$$\Upsilon_j(x) = \Phi^{-1}\left\{G_j(x)\right\}, \qquad j = 1,\ldots,p.$$

The dependence within this model is determined by the correlation matrix

$$\Sigma = \begin{pmatrix} 1 & \rho_{1,2} & \cdots & \rho_{p,1} \\ \rho_{1,2} & 1 & \cdots & \rho_{p,2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p,1} & \rho_{p,2} & \cdots & 1 \end{pmatrix}.$$

3.3 The multivariate Gaussian tail model

Inference for the multivariate Gaussian tail model for annual maxima data is based on the estimation of

1. the marginal distribution G_j parameters μ_j, σ_j and ξ_j, for each $j = 1, ..., p$;
2. the correlation matrix Σ.

For modelling the dependence for all sites within one of our regions, we would need a Gaussian tail model with up to 25 dimensions.

It would be too expensive computationally to estimate all $p(p-1)/2$ correlation parameters, so we look for a simplification based on the distances between sites.

3.3 The multivariate Gaussian tail model

We model the dependence between sites i and j as

$$\rho_{i,j} = \exp\left\{-\lambda\left[d(i,j)\right]^{\delta}\right\}$$

for parameters $\lambda \in \mathbb{R}_+$ and $0 < \delta < 2$, and where $d(i,j)$ denotes the Euclidean distance between sites i and j.

This method was motivated by Diggle *et. al.* (1998), and was also used in a different context by Smith and Walshaw (2003).

The Gaussian tail model in this form can feasibly be employed for high–dimensional problems, and we are able to use it for each of our 9 regions, adopting a Bayesian approach to inference.

3.4 Hierarchical model for the UK

The nine regions were originally defined by Wrigley *et al.* (1984) and represent regions of different physiographic character, within which the rainfall climate is regarded as being fairly coherent (Fowler and Kilsby, 2003).

Here we use random effects for the regional parameters. For a specific location and time, the GEV location and scale parameters vary as a function of space and time covariates.

Conditional on the model parameters, observations are independent between regions.

Our parameter vector is

$$\left(\mu_{i,j,k},\ \sigma_{i,j,k},\ \xi_k,\ \lambda_k,\ \delta_k\right)$$

where $i; i = 1, \ldots, 40$, indexes time (year); $j; j = 1 \ldots, 204$, indexes site, and $k; k = 1, \ldots, 9$ indexes region.

Note that the GEV location and scale parameters $\mu_{i,j,k}$ and $\sigma_{i,j,k}$ are only defined where site j falls in region k. The GEV shape parameters ξ_k are region–specific, as are the parameters λ_k and δ_k, which determine the correlation structure within region k.

3.4 Hierarchical model for the UK

For year i, site j and region k, we model the GEV parameters as follows:

$$\mu_{i,j,k} = \epsilon_{\mu_0}^{(k)} + \epsilon_{\mu_1}^{(k)} x^{(j)} + \beta_{\mu}^{(k)} i,$$

$$\eta_{i,j,k} = \log(\sigma_{i,j,k}) = \epsilon_{\sigma_0}^{(k)} + \epsilon_{\sigma_1}^{(k)} x^{(j)} + \beta_{\sigma}^{(k)} i$$

and

$$\xi_k = \epsilon_{\xi_0}^{(k)}.$$

Here $x^{(j)}$ is a site–specific covariate corresponding to the Daily Mean Rainfall. The location and log(scale) are thus varying as linear functions of this covariate, and time.

3.4 Hierarchical model for the UK

For two sites j_1 and j_2, we have

$$\rho^{(k)}_{j_1,\,j_2} = \exp\left[-\left\{\lambda^{(k)} d(j_1, j_2)\right\}^{\delta^{(k)}}\right],$$

where

$$\log(\lambda^{(k)}) = \epsilon^{(k)}_{\lambda}$$

and,

$$\log\left(\frac{2}{\delta^{(k)}} - 1\right) = \epsilon^{(k)}_{\delta}.$$

All of the terms

$$(\epsilon^{(k)}_{\mu_0}, \epsilon^{(k)}_{\mu_1}, \epsilon^{(k)}_{\sigma_0}, \epsilon^{(k)}_{\sigma_1}, \epsilon^{(k)}_{\xi}, \epsilon^{(k)}_{\lambda}, \epsilon^{(k)}_{\delta})$$

are random effects, and the next layer of our model is to specify distributions for these.

3.4 Hierarchical model for the UK: random effects

We follow Coles (1999) by specifying each ϵ_ω term to be Normally and independently distributed

$$\epsilon_\omega^{(k)} \sim N\left(a_\omega, \tau_\omega\right), \qquad k = 1, ..., 9$$

where, for convenience, we have specified the Normal distribution by its mean and precision.

The transformations on $\sigma_{i,j,k}$, $\lambda^{(k)}$ and $\delta^{(k)}$ ensure the correct ranges for these parameters.

3.4 Hierarchical model for the UK: priors

The bottom layer of the model is the specification of priors for the random effect distribution parameters. These are conjugate:

$$a_\omega \sim N(b_\omega, c_\omega), \qquad \tau_\omega \sim \Gamma(d_\omega, f_\omega)$$

with suitable choices for the hyperparameters $b_\omega, c_\omega, d_\omega$ and f_ω.

Here we treat the linear parameters as fixed effects (no relation between regions), and specify the following (non-conjugate) priors.

$$\beta_\omega^{(k)} \sim N(g_\omega, h_\omega), \qquad k = 1, ..., 9,$$

again for a suitable choice of hyperparameters g_ω and h_ω.

Inference is carried out via MCMC, where all random effects distribution parameters α_ω and τ_ω can be directly sampled from their posterior densities, while a Metropolis step is used for all other parameters.

Choices of the prior parameters were made to give rather un–informative, near flat prior distributions. Specifically we chose

$$b_\omega = 0, c_\omega = 10^{-6}, d_\omega = 10^{-1}, f_\omega = 10^{-3}, g_\omega = 0 \quad \text{and} \quad h_\omega = 10^{-6}.$$

The following results are based on 100,000 simulations based on 4 runs of length 35,000, with a burn-in discard of 10,000 each time. Units are $0.1mm$.

Region	ξ_0	λ^*	δ^*
CEE	-0.00896 (0.00324)	1.624 (0.083)	0.4906 (0.0859)
ES	-0.0201 (0.0046)	2.144 (0.105)	0.5712 (0.0405)
NEE	-0.00168 (0.00494)	1.497 (0.059)	0.5372 (0.0311)
NI	-0.00312 (0.00318)	2.792 (0.194)	0.6856 (0.0425)
NS	-0.0325 (0.0039)	2.087 (0.126)	1.015 (0.0627)
NWE	-0.00399 (0.00513)	2.370 (0.151)	0.7290 (0.0498)
SEE	-0.00612 (0.00316)	1.701 (0.097)	0.2470 (0.0643)
SS	-0.0345 (0.0023)	1.987 (0.090)	0.5939 (0.0440)
SWE	0.0155 (0.0043)	2.619 (0.269)	0.8476 (0.100)

Posterior means and standard deviations for the regional parameters

3.5 Inference and results

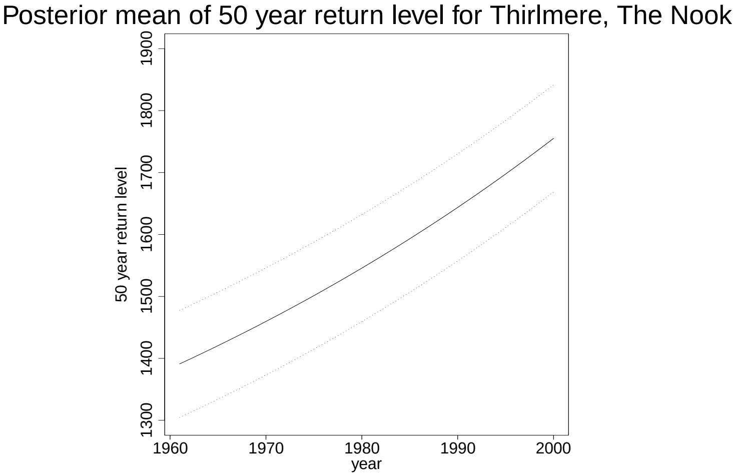
Region	β_μ	s.d.	β_σ	s.d.
CEE	1.404	(0.068)	0.02745	(0.00080)
ES	2.073	(0.075)	0.003762	(0.00066)
NEE	0.4632	(0.0738)	0.001036	(0.000767)
NI	0.2343	(0.0578)	0.01656	(0.00075)
NS	0.07580	(0.09294)	0.01069	(0.00077)
NWE	-0.1927	(0.0809)	0.01265	(0.00090)
SEE	-3.927	(0.136)	0.007268	(0.000950)
SS	-0.1097	(0.0665)	0.01829	(0.00067)
SWE	0.9108	(0.0739)	0.0003564	(0.0007531)

Posterior means and standard deviations for the linear trend parameters

3.5 Inference and results

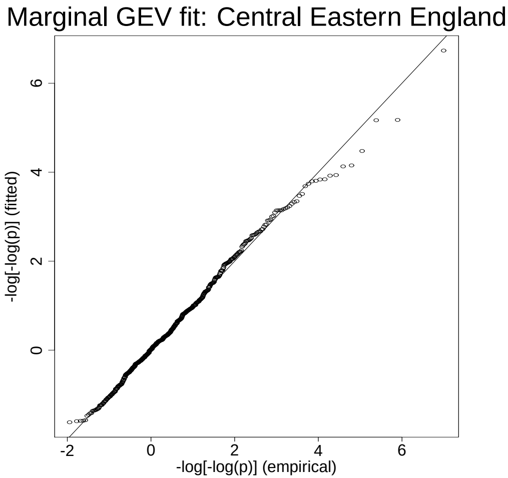
Site	μ	$\log(\sigma)$	ξ	50 year return level
Cambridge Niab	298.19 (1.90)	4.4834 (0.0211)	-0.00896 (0.00324)	637.76 (8.66)
Thirlmere, The Nook	800.98 (7.03)	5.5071 (0.0382)	-0.00399 (0.00513)	1755.9 (43.2)

Year 2000 parameters and 50 year return levels for two contrasting sites



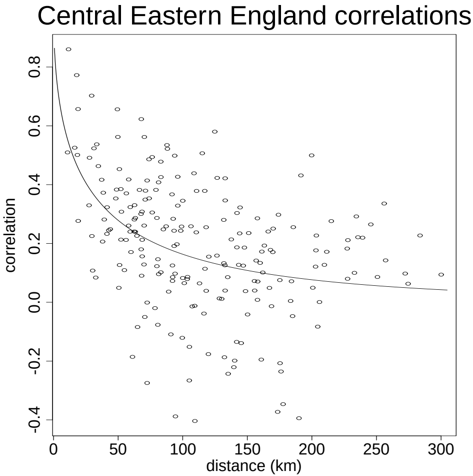
Posterior mean for 50 year return level (solid line) with 95% credibility bounds (dotted lines)

3.5 Inference and results



Marginal fitted versus empirical quantiles for individual sites, transformed to a $-log[-log(1-p)]$ scale, and then pooled

3.5 Inference and results



Empirical inter-site correlations (after transformation to Gaussian margins using marginal fits), plotted against distance separation. Solid line is model for dependence

References

Bortot, P., Coles, S. G. and Tawn, J. A. (2000) The multivariate Gaussian tail model: an application to oceanographic data. *Appl. Statist.*, **49**, 31–49.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998) Model-based geostatistics. *Appl. Statist.*, **47**, 299–350.

Fawcett, L. and Walshaw, D. (2008). Bayesian inference for clustered extremes. *Extremes*, 11, 217–233.

Fawcett, L. and Walshaw, D. (2006). Markov chain models for extreme wind speeds. *Environmetrics*, **17**, 795–809.

Fawcett, L. and Walshaw, D. (2006a). A hierarchical model for extreme wind speeds. *Appl. Statist.*, **55**, 5, 631–646.

Fowler, H. J. and Kilsby, C. G. (2003) A regional frequency analysis of United Kingdom extreme rainfall from 1961 to 2000. *Int. J. Climatol.*, **23**, 1313–1334.

Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Series*. Springer–Verlag, New York.

Smith, E. L. and Walshaw, D. (2003) Modelling bivariate extremes in a region. *Bayesian Statistics*, **7**, 681–690.

Wrigley, T. M. L., Lough, J. M. and Jones, P. D. (1984) Spatial patterns of precipitation in England and Wales and a revised, homogeneous England and Wales precipitation series. *J. Climatol.*, **4**, 1–25.