A Hierarchical Model for Extreme Wind Speeds

Lee Fawcett and Dave Walshaw

Newcastle University, Newcastle-upon-Tyne, U.K.

lee.fawcett@ncl.ac.uk

19th TIES Conference: Kelowna, British Columbia, June 2008

Structure of this talk

1. **The data**
2. **Model formulation**
   - Threshold exceedances
   - Site and seasonal variation
   - Temporal dependence
   - Model construction
3. **Analysis of the wind speed data**
   - MCMC
   - Model diagnostics
   - Return level inference

The data

In this talk we develop a hierarchical model for hourly maximum wind speeds over a region of central and northern England.

The data consist of hourly gust maximum wind speeds recorded for the British Meteorological Office at twelve locations (see Figure 1).

Data were collected hourly, over a period of eighteen years, giving about 157,000 observations per site.

The data

The sites used represent a variety of geographical locations:

- Both urban and rural
- Both high and low altitudes
- Easterly/westerly positions

Figure 2 illustrates an exploratory analysis of data from two contrasting sites, Nottingham and Bradfield.

The data

Figure: 1: Location of wind speed stations

The data

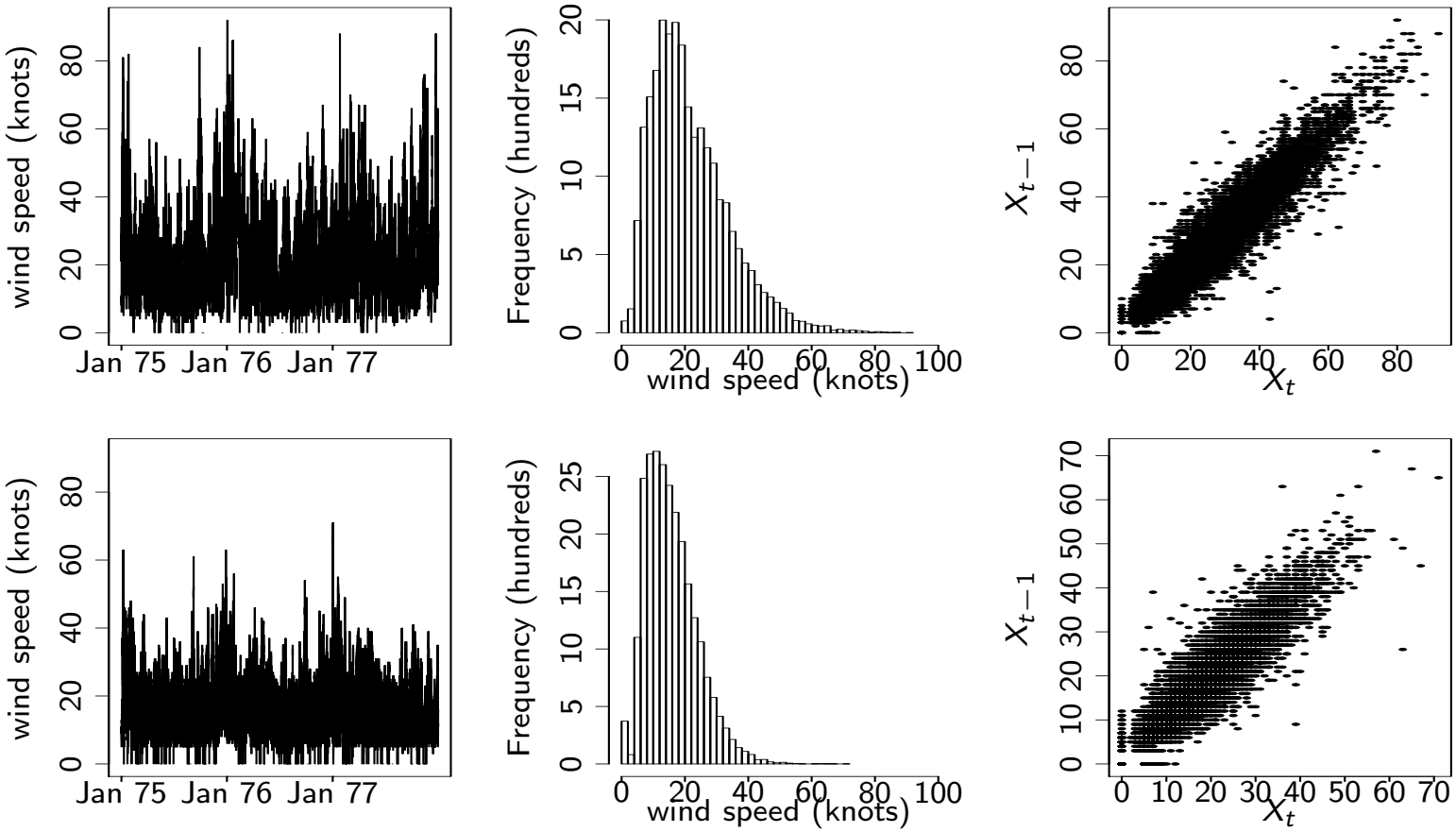

Figure: 2: Exploratory analysis of wind speed data

The data

In this talk, we construct a model which is based on a standard limiting extreme value distribution, but incorporates random effects for:

- variation across sites;
- seasonal variation, and
- the serial dependence

inherent in the time series of hourly maximum speed.

The data

The aims of this work are:

- To exploit the **complex structure** inherent in the data to improve over simplistic inferential procedures
- To build on the techniques used by **Coles (2002)** by
  - using a threshold–based approach to modelling
  - properly accounting for seasonal variation
  - explicitly modelling any temporal dependence in extreme wind speeds
- To adopt a Bayesian approach to inference and obtain **predictive return level estimates**

Modelling extremes

"Block maxima" approach

- e.g. *annual maximum wind speeds* might be modelled by an appropriate limiting distribution, such as the **generalised extreme value distribution**
- Highly inefficient!

Threshold methods

- An observation is extreme if it exceeds some high cut–off point (**threshold**)
- Use *all* observations which exceed this cut–off point – i.e. use *all* extremes!

Threshold methods

The generalised Pareto distribution (GPD)

Under very broad conditions, if it exists, any limiting distribution as $u \to \infty$ of $(X - u)|X > u$ is of GPD form, where

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

where $a_+ = \max(0, a)$ and σ $(\sigma > 0)$ and ξ $(-\infty < \xi < \infty)$ are **scale** and **shape** parameters (respectively).

Threshold methods

Practical implementation

1. Choose some threshold u which is high enough so that the GPD is a good model for $(X - u | X > u)$
2. Fit the GPD to the observed excesses $x - u$
3. Use the fitted GPD to provide estimates of extreme quantiles, or **return levels** (*see later*)

Seasonal variation

Possible solution: Restrict an extreme value analysis to the 'season' which contains the 'most extreme' extremes (e.g. **Coles and Tawn, 1991**)

We want our model to **take account** of seasonal variability and identify all gusts which are large *given the time of year* as extreme!

Our solution: Fit a seasonally–varying GPD!

- For wind speed data, there is no natural partition into separate seasons (in the UK)
- We partition the annual cycle into 12 'seasons' (we use calendar months)
  - reflects well the continuous nature of seasonal climate changes
  - still enough data within each season for analysis!

Site and seasonal variation

We take the same approach to allow for site variation.

Thus, our model will yield parameter pairs

$$(\sigma_{m,j}, \xi_{m,j}), \qquad \text{for } m, j = 1, \ldots, 12,$$

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

We also need our threshold u to vary, since different criteria for what constitutes an extreme will be in play for each combination of season and site.

We denote by $u_{m,j}$ the threshold for identifying extremes in month m at site j.

Temporal dependence

The plot of the time series against the version at lag 1, for each site, shows the presence of substantial serial correlation between successive extremes.

What can be done?

1. '**Remove**' it – use "Peaks over threshold" (**Davison and Smith, 1990**)
2. '**Ignore**' it – initially, but then adjust standard errors post–analysis (**Smith, 1991**)
3. **Model** it

Temporal dependence

The peaks over threshold (POT) approach is the most common tool used here. However,

- As with the block maxima approach, this is very wasteful of (precious!) data
- **Fawcett and Walshaw (2007)** make a strong case against this:
  - induces bias in GPD parameter estimation
  - results in underestimated return levels for processes with strong short–term temporal dependence

Temporal dependence

That leaves us with:

(a) Adjust inference post–analysis to account for temporal dependence

(b) Explicitly model the dependence present

We opt for (b), because:

- Our intention in this work is to investigate complex structures in the data, not *ignore* or *remove* them!
- Exploratory analyses support a simple first–order Markov model for the serial dependence (**Fawcett and Walshaw, 2006**)

First–order Markov structure

The stochastic properties of a first–order Markov chain are completely determined by the joint distribution of consecutive pairs.

Given a model $f(x_i, x_{i+1}; \boldsymbol{\psi})$ specified by parameter vector $\boldsymbol{\psi}$, the likelihood for $\boldsymbol{\psi}$ is given by

$$L(\boldsymbol{\psi}) \quad = \quad f(x_1; \boldsymbol{\psi}) \prod_{i=1}^{n-1} f(x_i, x_{i+1}; \boldsymbol{\psi}) \Big/ \prod_{i=1}^{n-1} f(x_i; \boldsymbol{\psi}).$$

Contributions to the numerator in the above can be modelled by using an appropriate **bivariate extreme value** model.

First–order Markov structure

The **logistic model** is one of the most flexible and accessible of these models (for example, see **Tawn, 1988**).

For consecutive threshold exceedances, the appropriate form of this model is given by:

$$F(x_i, x_{i+1}) \quad = \quad 1 - \left(Z(x_i)^{-1/\alpha} + Z(x_{i+1})^{-1/\alpha} \right)^{\alpha}, \qquad x_i, x_{i+1} > u,$$

where the transformation Z is given by

$$Z(x) \quad = \quad \lambda^{-1} \{ 1 + \xi(x-u)/\sigma \}_{+}^{1/\xi},$$

and ensures that the margins are of GPD form.

Independence and **complete dependence** are obtained when $\alpha = 1$ and $\alpha \searrow 0$ respectively.

First–order Markov structure

Inference is complicated by the fact that a bivariate pair may exceed a specified threshold in just one of its components.

Let

$$\begin{aligned} R_{0,0} &= (0,u)\times(0,u)\,, \\ R_{1,0} &= [u,\infty)\times(0,u)\,, \\ R_{0,1} &= (0,u)\times[u,\infty) \qquad \text{and} \\ R_{1,1} &= [u,\infty)\times[u,\infty)\,. \end{aligned}$$

For example, a point $(x_i, x_{i+1}) \in R_{1,0}$ if x_i exceeds the threshold but x_{i+1} does not.

First–order Markov structure

Inference

- The logistic model applies to points in $R_{1,1}$
- For points in $R_{1,0}$ or $R_{0,1}$, we use

$$\left.\frac{\partial F}{\partial x_i}\right|_{(x_i,u)} \quad \text{or} \quad \left.\frac{\partial F}{\partial x_{i+1}}\right|_{(u,x_{i+1})}$$

  respectively.
- For $R_{0,0}$, the contribution to the numerator in the Markov chain likelihood is given by the distribution function evaluated at the threshold u.

Assumptions

In the construction of our hierarchical model, we assume that:

- the GPD is valid for exceedances over a high threshold for each season at each site;
- extremes between sites and between seasons are independent,
  - but successive extremes *within* seasons have a first–order Markov dependence
  - independence *between* seasons seems reasonable – dependence between wind speed extremes is typically short–lived
  - Spatial dependence in weather at our sites should not translate to strong correlations between extremes across sites

Assumptions

We also assume that

- there is no interaction between seasonal and site effects
- both spatial effects and seasonal effects are exchangeable
  - we might expect there to be clear structure in the data – wind speeds in January and February should be more similar than those in January and July
  - we will return to this issue later on

Threshold stability

Recall that we denote by $(\sigma_{m,j}, \xi_{m,j})$ the parameters of the GPD assumed to be valid for threshold excesses in season m and site j.

To ensure **threshold stability** in our models, we now use

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

Threshold stability

With this parameterisation, if $X - u^*_{m,j}$ follows a GPD $(\tilde{\sigma}_{m,j}, \xi_{m,j})$, where $u_{m,j} > u^*_{m,j}$, then

- $X - u_{m,j}$ also follows the same GPD,
- which is useful for comparisons across different sites and seasons.
- It also allows us to specify prior information about both parameters without having to worry about threshold dependency.

Random effects model

With these assumptions in mind, we build the following **random effects model**:

$$\begin{aligned} \log(\tilde{\sigma}_{m,j}) &= \gamma_{\tilde{\sigma}}^{(m)} + \epsilon_{\tilde{\sigma}}^{(j)}, \\ \xi_{m,j} &= \gamma_{\xi}^{(m)} + \epsilon_{\xi}^{(j)} \qquad \text{and} \\ \alpha_j &= \epsilon_{\alpha}^{(j)}, \end{aligned}$$

where γ and ϵ represent **seasonal** and **site** effects respectively.

We work with $\log(\tilde{\sigma}_{m,j})$ for computational convenience, and to retain the positivity of the scale parameter $\tilde{\sigma}_{m,j}$.

Random effects model

All random effects for $\log(\tilde{\sigma}_{m,j})$ and $\xi_{m,j}$ are taken to be normally and independently distributed:

$$\begin{aligned} \gamma_{\tilde{\sigma}}^{(m)} &\sim N_0(0, \tau_{\tilde{\sigma}}) \qquad \text{and} \\ \gamma_{\xi}^{(m)} &\sim N_0(0, \tau_{\xi}), \qquad m = 1, \ldots, 12, \end{aligned}$$

for the **seasonal** effects, and

$$\begin{aligned} \epsilon_{\tilde{\sigma}}^{(j)} &\sim N_0(a_{\tilde{\sigma}}, \zeta_{\tilde{\sigma}}) \qquad \text{and} \\ \epsilon_{\xi}^{(j)} &\sim N_0(a_{\xi}, \zeta_{\xi}), \qquad j = 1, \ldots, 12, \end{aligned}$$

for the **site** effects.

In the absence of any prior knowledge about α_j, we set

$$\epsilon_{\alpha}^{(j)} \sim U(0, 1).$$

Random effects model

The final layer of the model is the specification of prior distributions for the random effect distribution parameters.

Here we adopt conjugacy wherever possible to simplify computations, specifying:

$$a_{\tilde{\sigma}} \sim N_0(b_{\tilde{\sigma}}, c_{\tilde{\sigma}}), \qquad a_{\xi} \sim N_0(b_{\xi}, c_{\xi});$$
$$\tau_{\tilde{\sigma}} \sim Ga(d_{\tilde{\sigma}}, e_{\tilde{\sigma}}), \qquad \tau_{\xi} \sim Ga(d_{\xi}, e_{\xi});$$
$$\zeta_{\tilde{\sigma}} \sim Ga(f_{\tilde{\sigma}}, g_{\tilde{\sigma}}), \qquad \zeta_{\xi} \sim Ga(f_{\xi}, g_{\xi}).$$

MCMC technique

Estimation of the model outlined in Section 2 is made via a **Metropolis within Gibbs** algorithm

- Here, we update each component singly using a Gibbs sampler where conjugacy allows;
- Elsewhere, we adopt a Metropolis step

MCMC technique

The **full conditionals** for the Gibbs sampling are:

$$
\begin{aligned}
a_\bullet|\ldots &\sim N\left(\frac{b_\bullet c_\bullet + \zeta_\bullet \sum \epsilon_\bullet^{(j)}}{c_\bullet + n_s\zeta_\bullet}, c_\bullet + n_s\zeta_\bullet\right);\\
\zeta_\bullet|\ldots &\sim Ga\left(f_\bullet + \frac{n_s}{2}, g_\bullet + \frac{1}{2}\sum(\epsilon_\bullet^{(j)} - a_\bullet)^2\right);\\
\tau_\bullet|\ldots &\sim Ga\left(d_\bullet + \frac{n_m}{2}, e_\bullet + \frac{1}{2}\sum(\gamma_\bullet^{(m)})^2\right);
\end{aligned}
$$

where $n_m = 12$ and $n_s = 12$.

The complexity of the GPD likelihood means that conjugacy is unattainable for the random effects parameters, and a Metropolis step is used here.

MCMC technique

Obviously, we first need to specify appropriate hyper–parameters. In the absence of any expert prior knowledge, we use:

$$b_{\cdot} = 0, \qquad c_{\cdot} = 10^{-6}, \qquad d_{\cdot} = e_{\cdot} = f_{\cdot} = g_{\cdot} = 10^{-2}.$$

The implementation of the MCMC scheme then yields samples from the approximate posterior distributions for

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

Some results

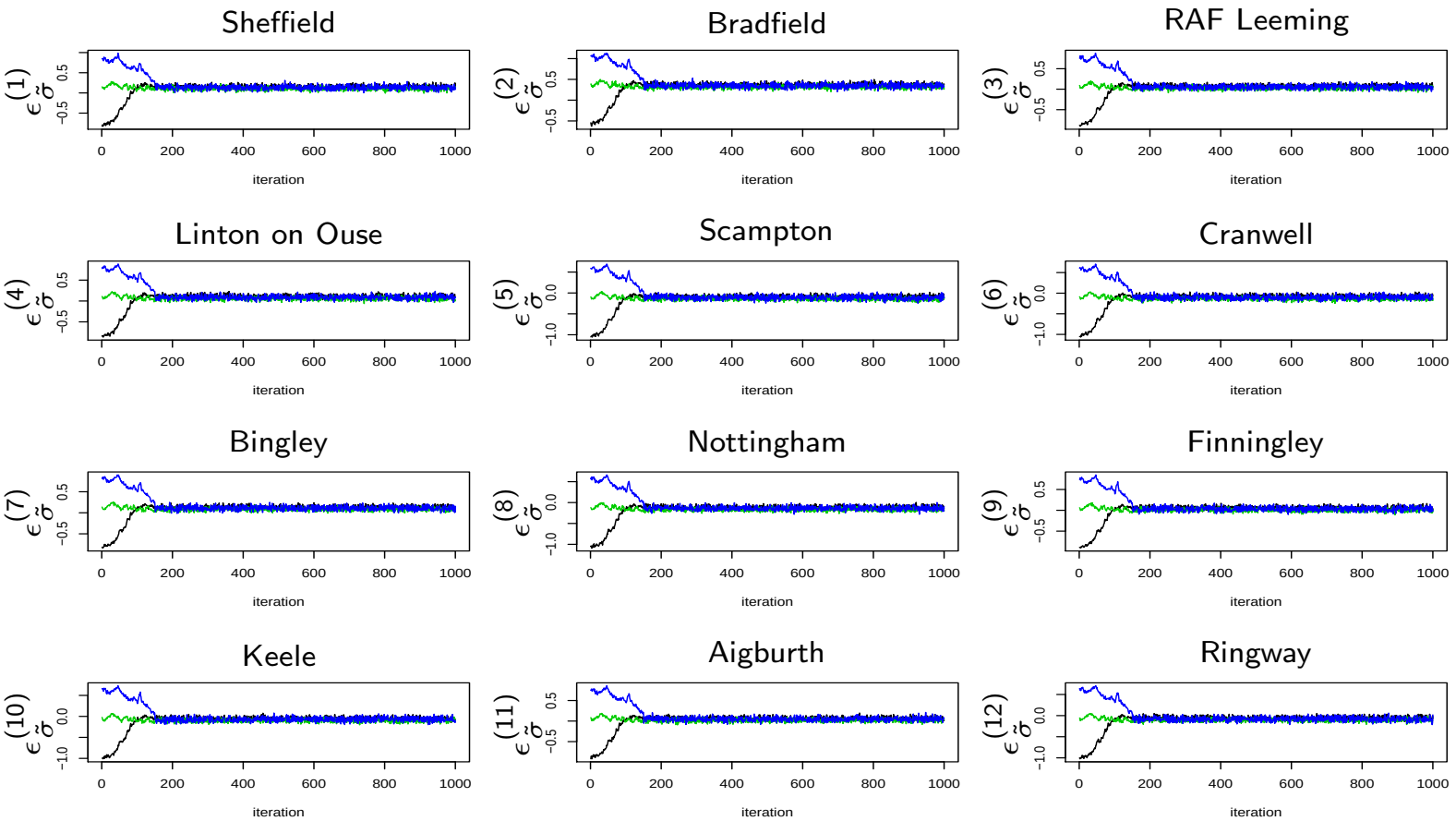

Figure: 3: MCMC output for site effects for $\log(\tilde{\sigma})$

Some results

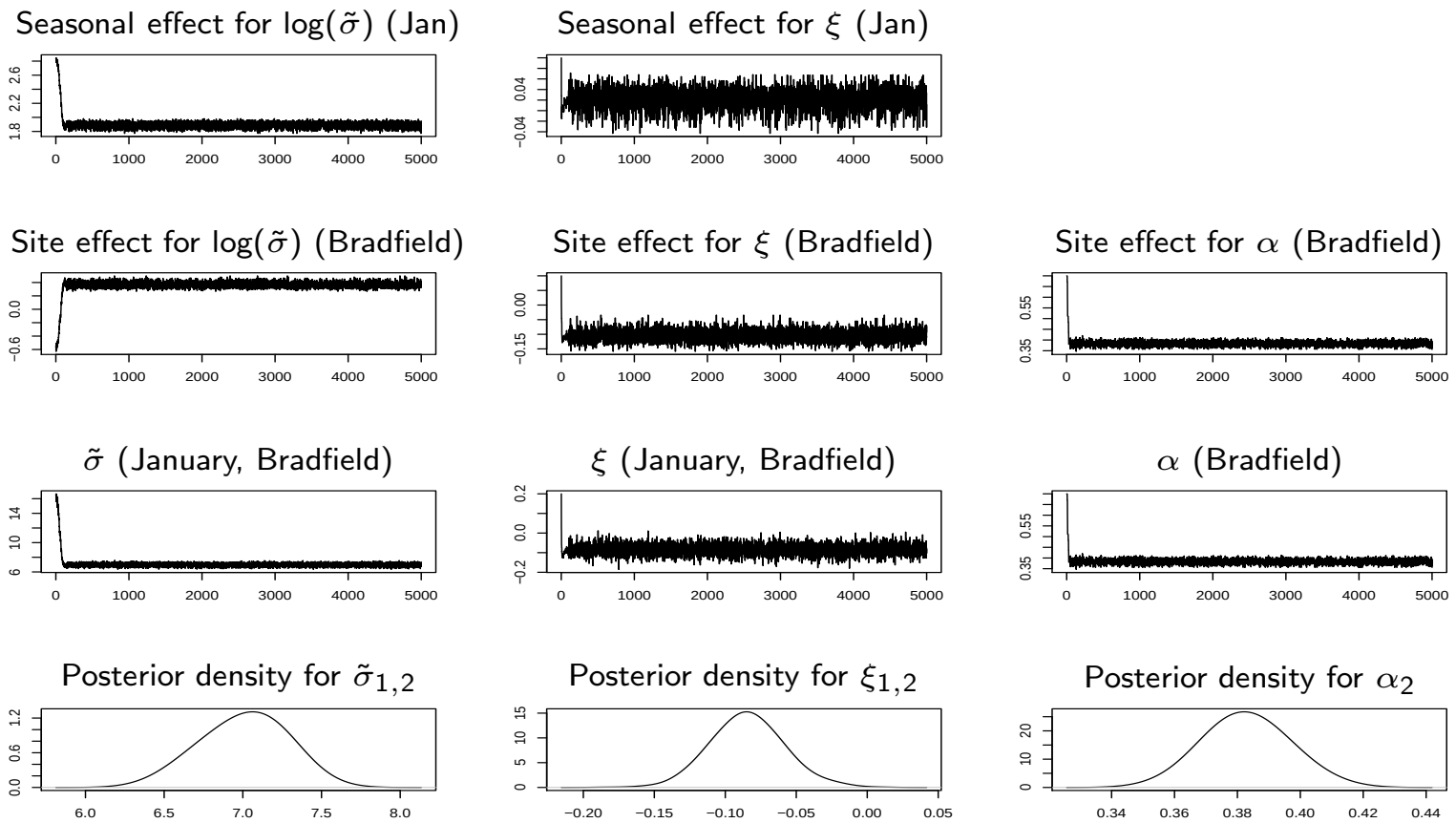

Figure: 4: MCMC output for Bradfield in January

Some results

	Bradfield, January	**Nottingham, July**
	Mean (st. dev.) MLE (s.e.)	Mean (st. dev.) MLE (s.e.)
$\gamma_{\tilde{\sigma}}^{(m)}$	1.891 (0.042)	1.294 (0.042)
$\gamma_{\xi}^{(m)}$	0.021 (0.018)	0.002 (0.018)
$\epsilon_{\tilde{\sigma}}^{(j)}$	0.367 (0.044)	–0.121 (0.041)
$\epsilon_{\xi}^{(j)}$	–0.105 (0.020)	–0.059 (0.017)
$\epsilon_{\alpha}^{(j)}$	0.385 (0.009)	0.300 (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)

Table: 1: Bayesian random effects analysis

Return level inference

Let $X_1, X_2, \ldots, X_n$ be the first n observations from a stationary sequence with marginal distribution function F.

Standard arguments in **Leadbetter et al. (1983)** show that, for large n and x,

$$\Pr\{\max(X_1, X_2, \ldots, X_n) \leq x\} \approx \{F(x)\}^{n\theta},$$

where $\theta \in (0, 1]$ is the **extremal index** and is a measure of the degree of extremal dependence in the series.

Setting $x = q_r$ in the above extression, equating this to $1 - r^{-1}$ and solving for q_r, gives, to a good approximation, the **r–year return level**.

Accurate and precise return level estimation is an important design consideration.

Return level inference

For each site j, $j = 1, \ldots, 12$, the annual exceedance rate of q_r is given by

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

where

- $\left\{1 - F_{m,j}\left(q_r\right)^{h_{m,j}\theta_j}\right\}$ 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.

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

Table: 2: Return levels for Bradfield (knots)

Shrinkage plots

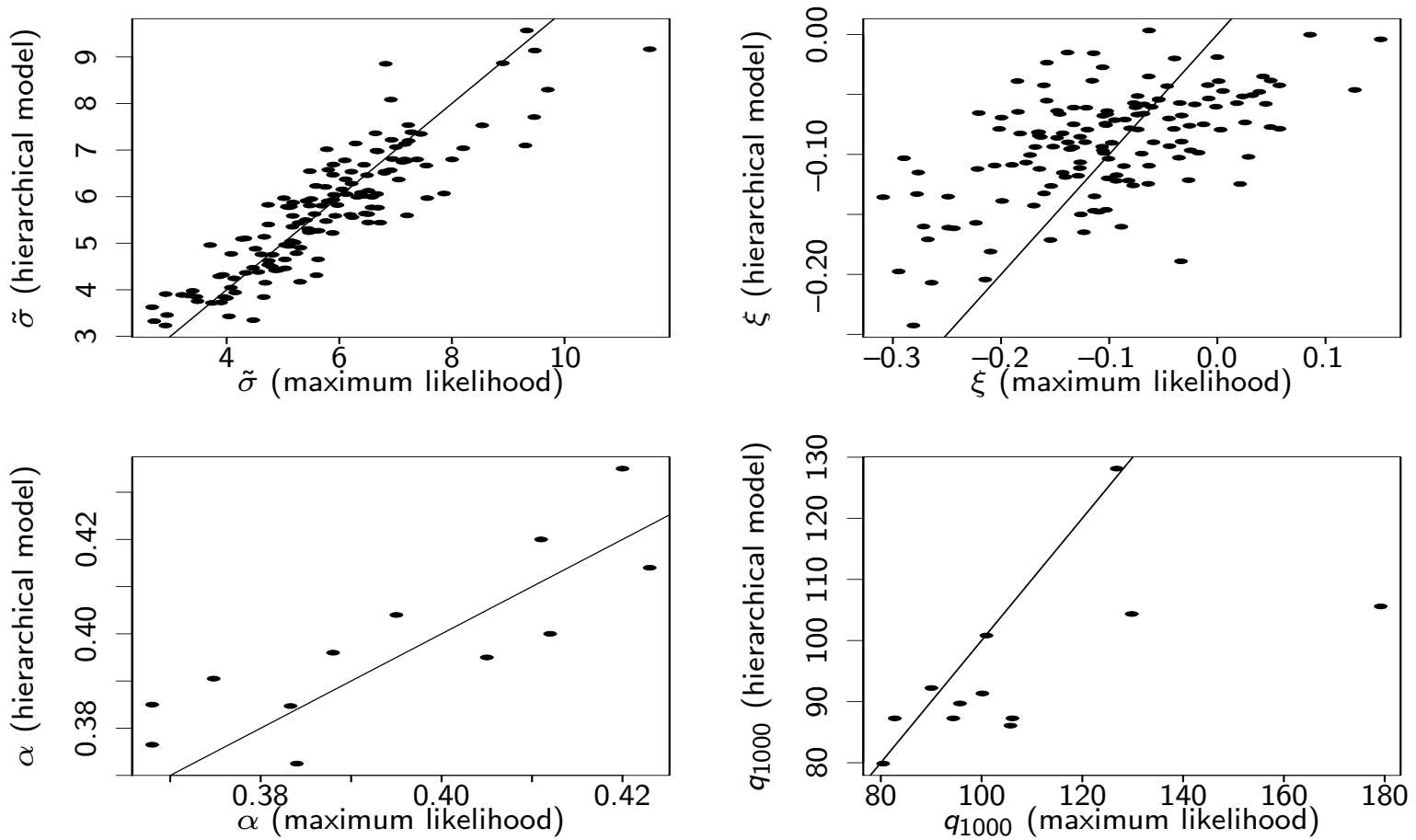

Figure: 5: Posterior means against maximum likelihood estimates

Summary of results

The main take–home points are:

- A reduction in sampling variation under the Bayesian hierarchical model
  - posterior standard deviations substantially smaller than the corresponding standard errors...
  - ... probably due to the pooling of information across sites and seasons
  - This is also evident in the "shrinkage plots"
- Estimates of return levels using maximum likelihood estimation can be very unstable – the hierarchical model achieves a greater degree of stability
- The Bayesian paradigm offers an extension to **predictive return levels**, which cannot be achieved under the classical approach to inference

References

Coles, S. G. (2002). A Random Effects Model in Extreme Value Analyses. Preprint.

Coles, S.G. and Tawn, J.A. (1991). Modelling extreme multivariate events. *J. R. Statist. Soc., B*, **53**, 377–392.

Davison, A.C. and Smith, R.L. (1990). Models for Exceedances over High Thresholds (with discussion). *J. R. Statist. Soc., B*, **52**, 393–442.

Fawcett, L. and Walshaw, D. (2007). Improved estimation for temporally clustered extremes. *Environmetrics*, **18**, 173–188.

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

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

Smith, R.L. (1991). Regional estimation from spatially dependent data. Preprint. (http:www.stat.unc.edu/postscript/rs/regest.pdf)

Tawn, J.A. (1988). Bivariate extreme value theory: Models and estimation. *Biometrika*, **75**, 397–415.