Modelling Environmental Extremes

Lee Fawcett and Dave Walshaw

School of Mathematics & Statistics

Newcastle University, UK

Short course for the 19th annual conference of

The International Environmetrics Society

Sunday, June 8th 2008

The University of British Columbia Okanagan,

Kelowna, Canada

1 Classical models and threshold models

1.1 Introduction

Statistical modelling of extreme weather has a very practical motivation: reliability — anything we build needs to have a good chance of surviving the weather/environment for the whole of its working life. This has obvious implications for civil engineers and planners. They need to know:

- how strong to make buildings;
- how high to build sea walls;
- how tall to build reservoir dams;
- how much fuel to stockpile;

etc.

This motivates the need to estimate what the:

- strongest wind;
- highest tide;
- heaviest rainfall;
- most severe cold-spell;

etc. will be over some fixed period of future time. The only sensible way to do this is to use data on the variable of interest (wind, rain etc.) and fit an appropriate statistical model. The models themselves are motivated by asymptotic theory, and this is our starting point.

1.2 Classical models

Extreme value modelling has a central theoretical result, analogous to the Central Limit Theorem. Suppose $X_1, X_2, \ldots,$ is an independent and identically distributed sequence of random variables. Define

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

We are interested in the limiting distribution of M_n as $n \to \infty$. As with the mean, $\bar{X}$, of $\{X_1, \ldots, X_n\}$, the limiting distribution of M_n as $n \to \infty$ is *degenerate*, and we need to work with a normalised version.

1.2.1 The Extremal Types Theorem (Fisher and Tippett, 1928)

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\Pr\{(M_n - b_n)/a_n \leq z\} \to G(z) \quad \text{as} \quad n \to \infty,$$

where G is a non–degenerate distribution function, then G belongs to one of the following families:

$$\begin{aligned} I : G(z) &= \exp\left\{-\exp\left[-\left(\frac{z-\beta}{\gamma}\right)\right]\right\}, \quad -\infty < z < \infty; \\ II : G(z) &= \exp\left\{-\left(\frac{z-\beta}{\gamma}\right)^{-\alpha}\right\}, \quad z > \beta; \quad [G(z) = 0, z \leq \beta]; \\ III : G(z) &= \exp\left\{-\left[-\left(\frac{z-\beta}{\gamma}\right)^{\alpha}\right]\right\}, \quad z < \beta; \quad [G(z) = 1, z \geq \beta], \end{aligned}$$

for parameters $\gamma > 0, \beta$, and $\alpha > 0$.

1.2.2 The Generalised Extreme Value Distribution (GEV)

Families I, II and III are widely referred to as Gumbel, Frechet and Weibull (or Extreme Value Types I, II and III) respectively.

Fortunately they can be combined into a single family, known as the Generalised Extreme Value Distribution (GEV), with c.d.f.

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

defined on the set $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$, and where μ, $\sigma > 0$ and ξ are *location*, *scale* and *shape* parameters respectively.

Note that the Extreme Value Types I, II and III correspond to the cases $\xi = 0$, $\xi > 0$ and $\xi < 0$ respectively.

For Type I, we need to take the limiting form of Equation (1) as $\xi \to 0$, which gives

$$G(z) = \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\}, \tag{2}$$

defined for all z.

So the Extremal Types Theorem can be restated with (1) as the limiting form, and this provides the basis for our first modelling approach.

Approach 1: "Block maxima"

Break up our sequence $X_1, X_2, \ldots$ into blocks of size n (with n reasonably large), and extract only the maximum observation from each block.

Now fit Model (1) to the sequence of extracted maxima $M_{(1)}, M_{(2)}, \ldots, M_{(N)}$ and use this as the basis for statistical inference. The most common implementation of this approach for weather data is to take our block size to be one year. This rough and ready approach has shown itself to be surprisingly robust!

1.2.3 Example: Annual maximum rainfall

Consider the annual maxima of daily rainfall accumulations (*mm*) at a location in SW England, from 1914 to 1961.

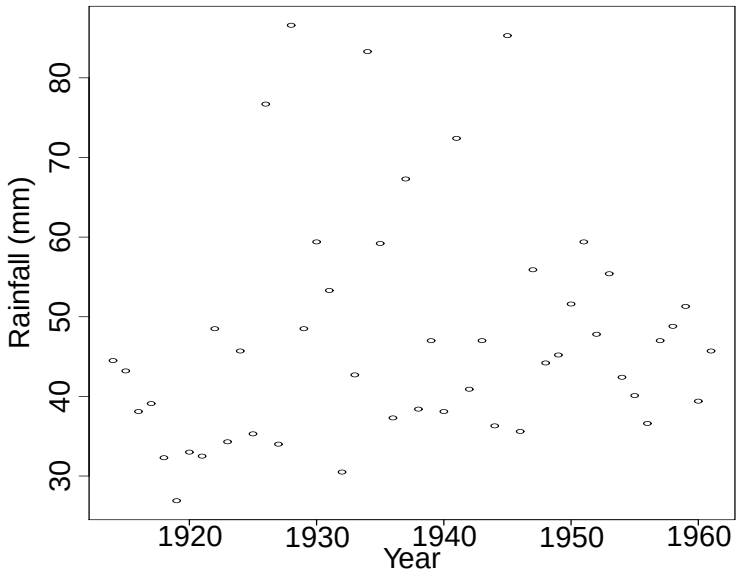

Figure 1: Annual maxima of daily rainfall totals at a location in South West England

1.2.4 Inferences for the block maxima approach

Here our blocks have $n = 365$, which is reasonably large, so we fit Model (1) to the $N = 48$ annual maxima (e.g. using *maximum likelihood estimation*). We obtain fitted parameter values (standard errors in parentheses):

$$\mu = 40.7(1.5) \qquad \sigma = 9.4(1.2) \qquad \xi = 0.14(0.12).$$

More importantly, we can make inferences on the quantities most useful to practitioners .... For example, the 99th percentile in the distribution of annual maxima is known as the *100 year return level*. The fitted value of this is easily obtained on inversion of Model (1):

$$q_{100} = 101.3(18.9).$$

1.2.5 Remarks about the block maxima approach

- We don't need to deal explicitly with normalisation constants. We don't even need to know n!
- The assumption of n independent and identically distributed variables in each block is cavalier, but inferences are surprisingly robust.
- The inferences on return levels are crucial for designers and engineers, to the extent they are built into legally binding codes of practice.

- In actual fact, the existing codes of practice are usually based on a very primitive version of the methods just described. Fits are often based on restricting to one of the Fisher–Tippett types, ignoring estimation uncertainty, and using an *ad hoc* interpolation of return levels across a network of sites.

- In any case the block–maxima approach is often *very* wasteful of data, leading to large uncertainties on return level estimates. This motivates a different approach . . . ...

1.2.6 Diagnostics for the block maxima approach

The goodness–of–fit of the GEV model is most easily assessed using various diagnostic plots. Here we consider four plots:

1. **Probability plot:** the fitted value of the c.d.f. is plotted against the empirical value of the c.d.f. for each data point.

2. **Quantile plot:** the empirical quantile is plotted against the fitted quantile for each data point.

3. **Return level plot:** the return level (with error bars) is plotted against the return period. Each data point defines a sample point.

4. **Density plot:** the fitted p.d.f. is superimposed on a histogram of the data.

For our rainfall example, the diagnostic plots look like this . . .

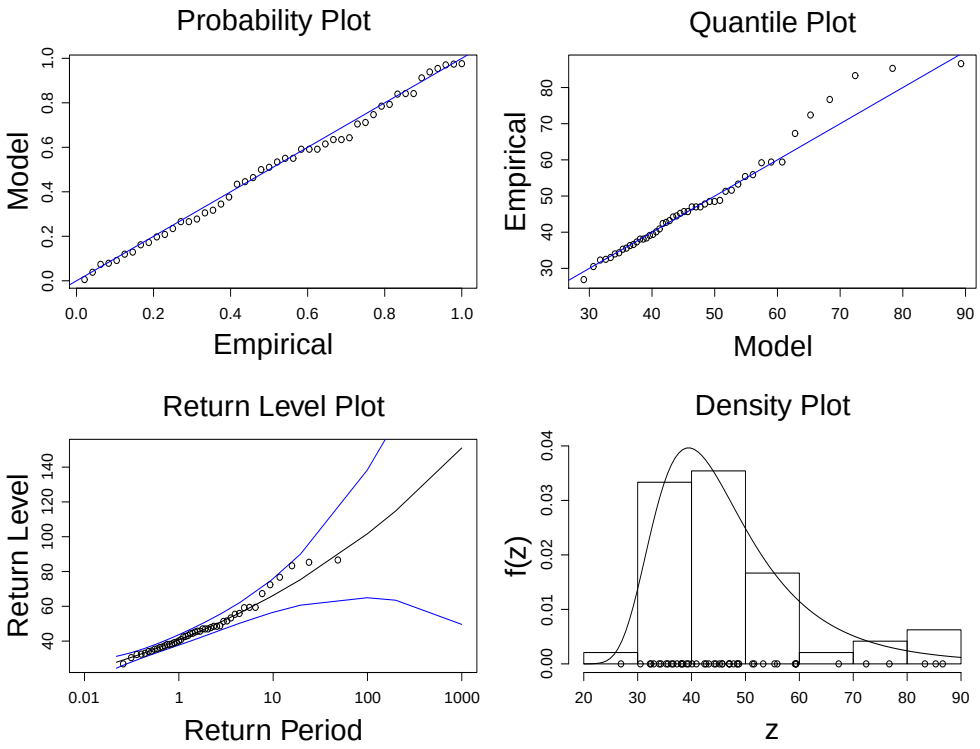

Figure 2: Diagnostic plots for GEV fit to rainfall annual maxima

1.2.7 Confidence intervals for return levels

Although we could construct a symmetrical confidence interval for the r–year return–level using classical likelihood theory ($\hat{q}_r \pm 1.96 \times$ standard error), this is not recommended. This practice assumes the limiting quadratic behaviour of the likelihood surface near the maximum, whereas in fact the surface is usually very asymmetrical.

We recommend using the method of *profile likelihood* to take this into account: by reparameterisation of Equation (1) to replace one of the parameters by q_r, we can maximise the likelihood *conditional* on q_r taking each possible value. We plot this constrained value against q_r ...

1.2.8 Profile likelihood confidence interval for q_{100}

For the rainfall example we get:

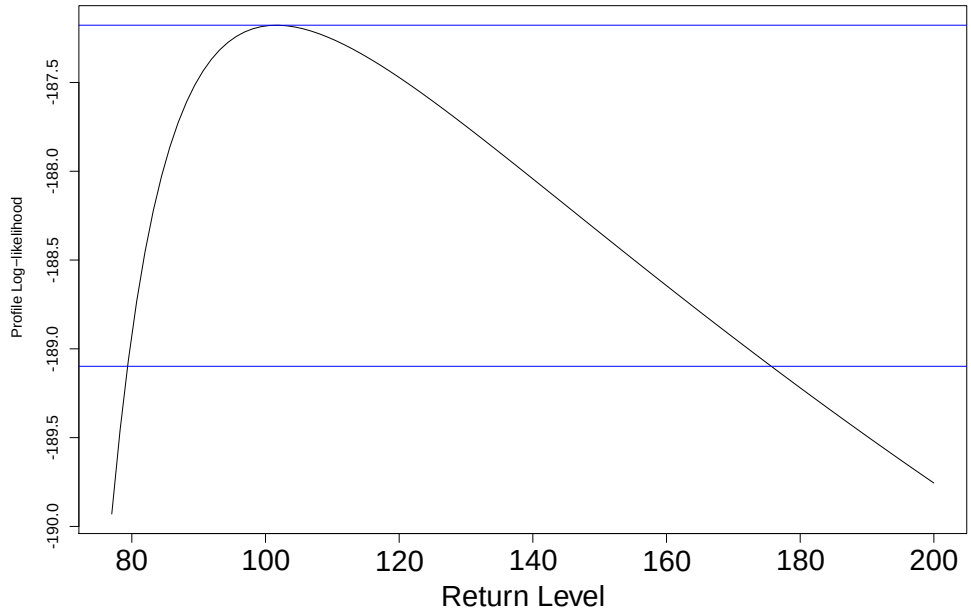

Figure 3: Profile log-likelihood for 100 year return level

The likelihood–ratio test can be applied directly to this likelihood surface by using a cut–off equal to $0.5 \times \chi^2_1(.)$. Here we see that the 95% confidence interval is approximately (78,176).

1.3 Threshold methods

Threshold methods use a more natural way of determining whether an observation is extreme - *all* values greater than some high value (*threshold*) are considered. This allows more efficient use of data, but brings its own problems. We must first go back and consider the asymptotic theory appropriate for this new situation.

1.3.1 The Generalised Pareto Distribution (GPD)

The appropriate limit theorem can be stated as follows:

Under very broad conditions, if it exists, any limiting distribution as $u \to \infty$ of $(X - u|X > u)$ is of Generalised Pareto Distribution (GPD) form (setting $Y = X - u$):

$$H(y) \;=\; 1 - \left(1 + \frac{\xi y}{\sigma}\right)_+^{-1/\xi}, \tag{3}$$

where $a_+ = \max(0, a)$ and σ ($\sigma > 0$) and ξ ($-\infty < \xi < \infty$) are scale and shape parameters respectively. Once again the GPD exists for $\xi = 0$, and is given by taking the limit of (3) as $\xi \to 0$. This time we get

$$H(y) \;=\; 1 - \exp\left(\frac{-y}{\sigma}\right), \tag{4}$$

defined for $y > 0$. This shows that when $\xi = 0$, the GPD is in fact the Exponential Distribution with mean equal to the scale parameter σ ($\sigma > 0$).

1.3.2 Return levels for the threshold excesses approach

If the GPD is a suitable model for exceedances of a threshold u by a random variable X, then for $x > u$,

$$\Pr\{X > x|X > u\} = \left[1 + \xi\left(\frac{x - u}{\sigma}\right)\right]^{-1/\xi}.$$

It follows that

$$\Pr\{X > x\} = \lambda_u \left[1 + \xi\left(\frac{x - u}{\sigma}\right)\right]^{-1/\xi}. \tag{5}$$

where $\lambda_u = \Pr\{X > u\}$. So the level x_m that is exceeded once every m observations is the solution of

$$\lambda_u \left[1 + \xi\left(\frac{x - u}{\sigma}\right)\right]^{-1/\xi} = \frac{1}{m}.$$

Rearranging this we obtain

$$x_m = u + \frac{\sigma}{\xi}[(m\lambda_u)^\xi - 1],$$

so long as m is large enough to ensure that $x_m > u$. Now if there are n_y observations per year, then by setting $m = N \times n_y$, the N–year return level is obtained as

$$z_N = \mu + \frac{\sigma}{\xi}[(Nn_y\lambda_u)^\xi - 1] \tag{6}$$

or when $\xi = 0$,

$$z_N = u + \sigma \log(Nn_y\lambda_u),$$

and standard errors can be obtained using the delta method.

Approach 2: "Exceedances over thresholds"

In practice, modelling might typically proceed as follows:

1. Choose some threshold u_0 which is high enough so that the GPD (3) is a good model for $(X - u_0|X > u_0)$.
2. Fit the GPD to the observed excesses $x - u_0$.
3. Use the fitted GPD, together with some model for the rate of exceedances $X > u_0$, to provide estimates for *return levels* using (6).

1.3.3 Example: daily rainfall totals

For the rainfall data we used before, now consider the daily totals themselves.

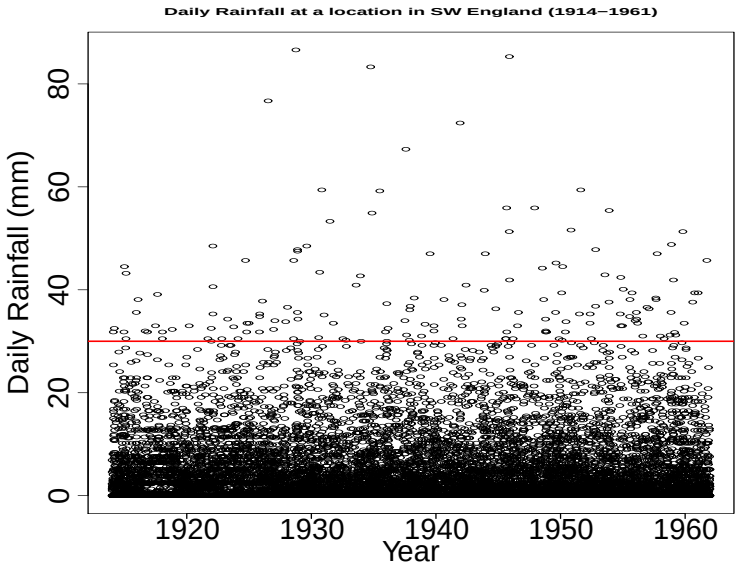

Figure 4: Daily Rainfall (1914-1961)

1.3.4 Threshold choice: Mean residual life plot

We make use of the fact that if the GPD is the correct model for all the exceedances x_i above some high threshold u_0, then the *mean excess*, i.e. the mean value of $(x_i - u)$, plotted against $u > u_0$, should give a linear plot (Davison and Smith, 1990) [Because $E[X_i - u_0]$ is a linear function of $u : u > u_0$]. By producing such a plot for values of u starting at zero, we can select reasonable candidate values for u_0.

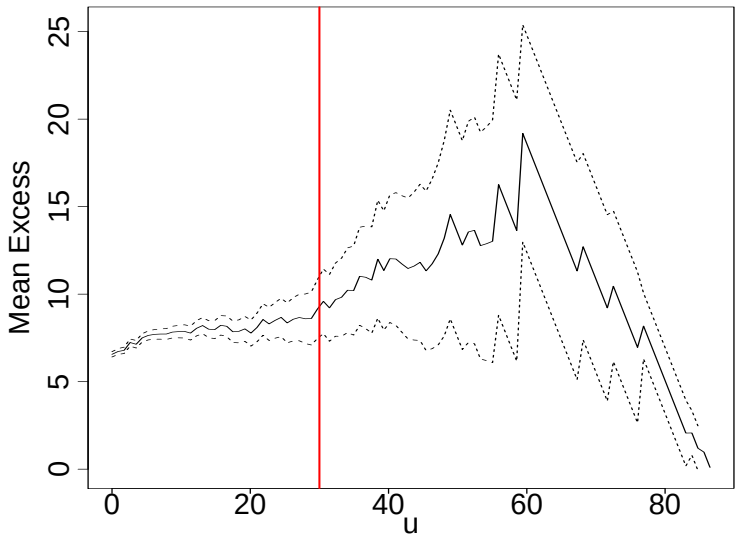

Figure 5: Mean residual life plot for daily rainfall

1.3.5 Inferences for the rainfall threshold excesses

Model (3) turns out to work reasonably well for all the excesses above $u_0 = 30mm$. This gives 152 exceedances $x_i; i = 1, \ldots, 152$, and Model (3) is fitted to the values $(x_i - u)$, again using maximum likelihood. We get

$$\sigma = 7.44(0.96) \qquad \xi = 0.18(0.10).$$

Assuming a uniform rate of exceedances, we estimate the 100–year return level: $q_{100} = 106.3(20.8)$.

1.3.6 Diagnostics for the rainfall threshold excesses

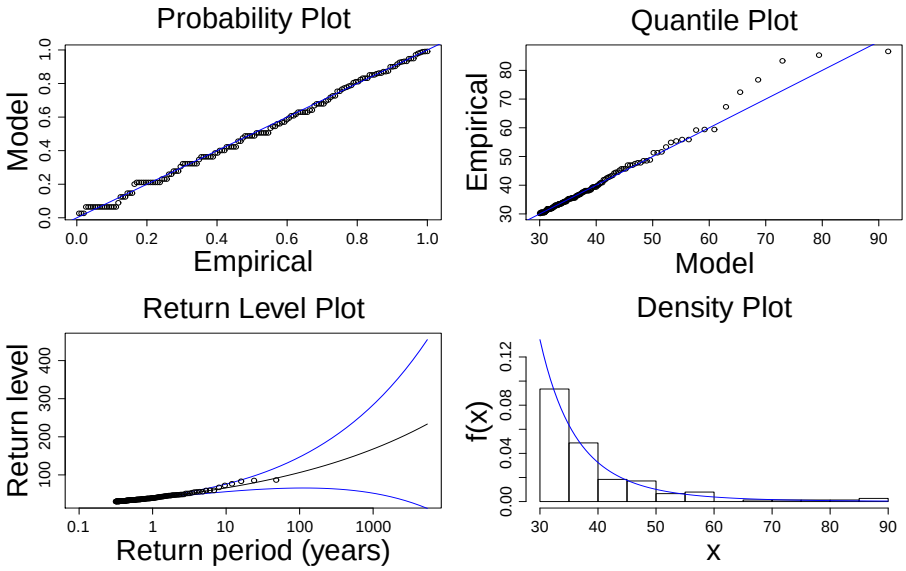

Figure 6: Diagnostic plots for the the threshold exceedance model for rainfall

1.3.7 Profile likelihood confidence interval for q_{100}

From the graph below, the 95% confidence interval is approximately (81,184).

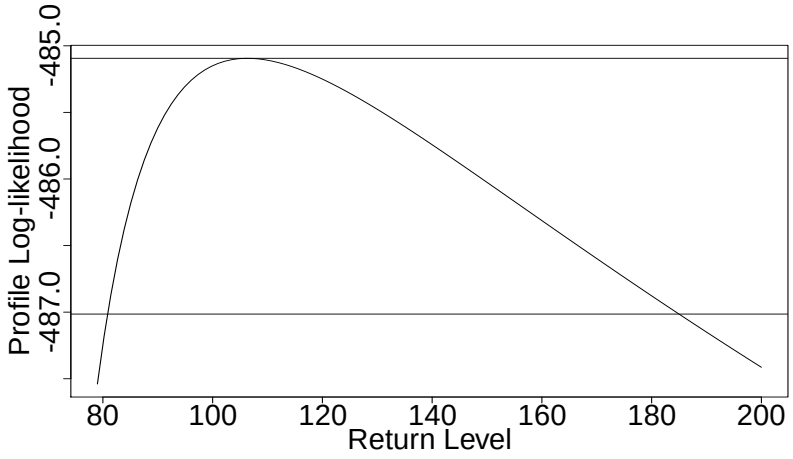

Figure 7: Profile log–likelihood for q_{100} based on threshold excess model

1.3.8 Threshold choice revisited

If the GPD with shape parameter ξ and scale parameter σ_{u_0} is the correct model for excesses over u_0, then for any threshold $u > u_0$, the excesses will be GPD with shape parameter ξ, and scale parameter

$$\sigma_u = \sigma_{u_0} + \xi(u - u_0).$$

If we now use a modified version of the scale parameter,

$$\sigma^* = \sigma_u - \xi u,$$

we can see that both σ^* and ξ should be constant over thresholds greater than u_0 if we model excesses $x_i - u$ for $u > u_0$ using the GPD. This provides us with a further tool for assessing our original choice of threshold u_0.

1.3.9 Parameter stability plots

We refit the GPD for a range of thresholds upwards of u_0, and investigate the stability of our estimates of ξ and σ^*. 95% confidence intervals are shown by vertical lines, and help us assess the significance of any variation we see.

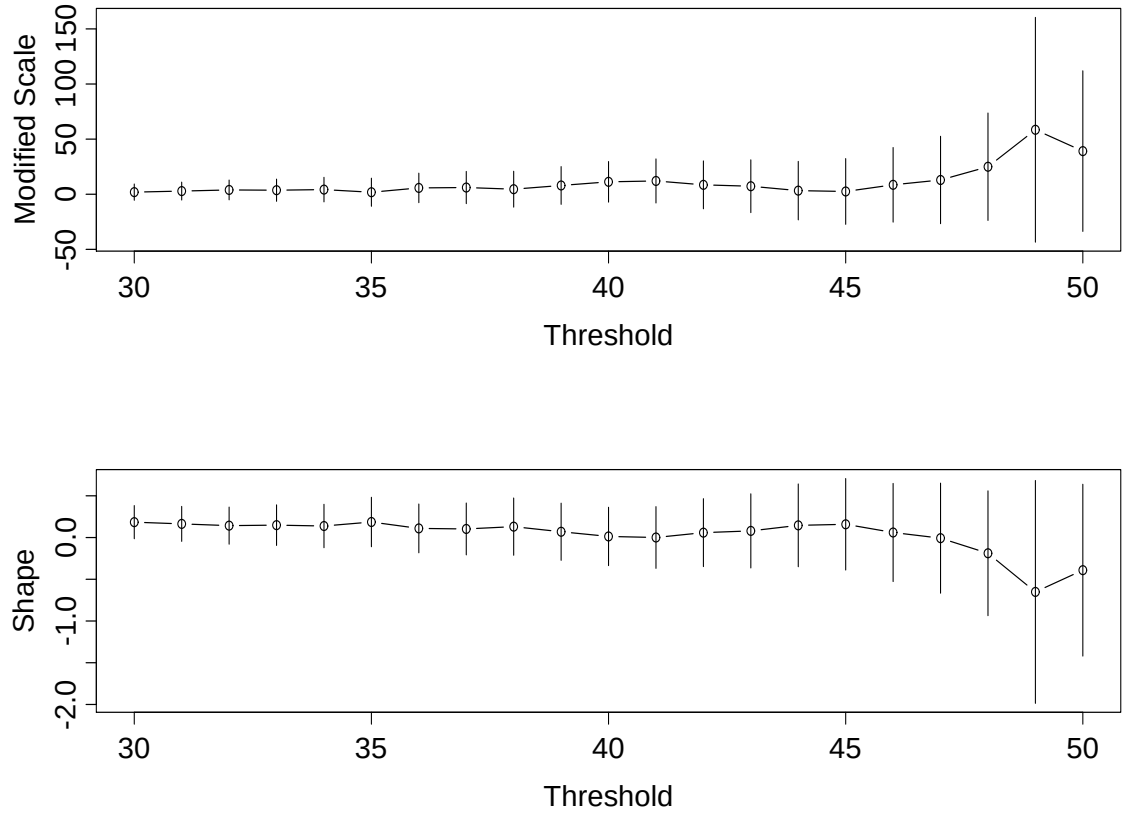

Figure 8: Parameter stability plots for the threshold model for rainfall

We can be reassured about our original choice of $u_0 = 30$!