Accounting for Choice of Measurement Scale in Extreme Value Modeling

J. L. Wadsworth*
Department of Mathematics and Statistics
Lancaster University, LA1 4YF, UK

J. A. Tawn
Department of Mathematics and Statistics
Lancaster University, LA1 4YF, UK

P. Jonathan
Shell Technology Centre Thornton
PO Box 1, Chester, UK

October 21, 2009

Abstract

We investigate the effect that the choice of measurement scale has upon inference and extrapolation in extreme value analysis. Separate analyses of variables from a single process on scales which are linked by a non-linear transformation may lead to discrepant conclusions concerning the tail behavior of the process. We propose the use of a Box-Cox power transformation incorporated as part of the inference procedure to account parametrically for the uncertainty surrounding the scale of extrapolation. This has the additional feature of increasing the rate of convergence of the distribution tails to an extreme value form in certain cases and thus reducing bias in the model estimation. Inference without reparameterization is practicably infeasible, so we explore a reparameterization which exploits asymptotic theory of normalizing constants required for non-degenerate limit distributions. Inference is carried out in a Bayesian setting via MCMC, an advantage of this being the availability of posterior predictive return levels. The methodology is illustrated on both simulated data and significant wave height data from the North Sea.

1 Introduction

The usual objective of extreme value analysis is to use sample data from rare events of a process to make rational predictions about the likely levels of future extremes of the process. To do this one models extreme data using an asymptotically justified probability model. The most fundamental such example is the Generalized Extreme Value (GEV) distribution. The GEV arises as the limiting law for appropriately normalized maxima of a wide variety of random variables; it is a three parameter distribution with distribution function

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

where μ, σ, ξ are respectively location, scale and shape parameters, and $z_+ = \max\{0, z\}$. This distribution is herein denoted GEV(μ, σ, ξ). The cases $\xi > 0$, $\xi = 0$ (interpreted as the limit $\xi \to 0$) and $\xi < 0$ are sometimes referred to as the Fréchet, Gumbel and Negative Weibull types, respectively. Other approaches to modeling extreme data are discussed in Section 3. Mathematical details of univariate extreme value theory are covered extensively in Leadbetter *et al.* (1983), whilst more statistical aspects are treated in e.g. Coles (2001).

There are many applications of extreme value analysis where data pertaining to the same physical process may naturally be measured on more than one scale. If the transformation between measurement scales

*j.wadsworth@lancaster.ac.uk

is linear, the appropriate type of extreme value distribution remains unaltered. If, on the other hand, a non-linear transformation is applied, different limiting distributions may be appropriate. Applying extreme value methods to the data on these different scales can lead to disparate conclusions regarding future extremes. This paper proposes methodology which allows the modeler to take into account their uncertainty over the scale upon which to conduct extreme value analysis.

As a motivating example consider the following. In ocean engineering, significant wave height (H_s), defined as four times the standard deviation of displacement from mean sea level, is a measure of ocean energy. Understanding of the extremes of this variable are vital for offshore structural design. However, one might equally wish to consider the extremes of the drag force induced by the waves on a fixed offshore structure, a variable which is proportional to the square of H_s (Tromans and Vanderschuren, 1995). Although the two variables are measurements of same physical process, differing conclusions may be derived concerning their tail behavior. For the wave height data to be considered in Section 4, a simple likelihood-based analysis of weekly maxima of H_s produces a 100-year return level estimate and 95% confidence interval of 14.66 meters (13.63, 16.35). However analysing H_s^2 instead, then back-transforming the results to the H_s scale, the estimate becomes 16.27 meters (14.51, 18.92). Furthermore the estimated shape parameters of the two variables differ markedly: for H_s, $\hat{\xi} = -0.12$ $(-0.17, -0.06)$, whereas for H_s^2, $\hat{\xi} = 0.11$ $(0.04, 0.19)$. These results suggest light-tailed behavior with a finite upper end point for H_s, yet heavy-tailed behavior with no finite upper end point for H_s^2. Such a situation gives rise to increasingly discrepant return level inferences with lengthening return period. It seems natural therefore to account for this uncertainty over the scale on which to extrapolate as part of the inference.

We approach this problem by incorporating a power transformation into the inference procedure; specifically we use the well-known Box-Cox transformation (Box and Cox, 1964). This transformation offers the possibility of improving the rate of convergence to the limiting extreme value form, since different distributions converge at different rates. This type of transformation restricts the methodology to cases where the extreme data are strictly positive, however this encompasses a wide variety of practical problems. The use of the Box-Cox transformation in extreme value analysis has been considered before in an entirely different context in the work of Eastoe and Tawn (2009). In their work the motivation was the standardization of non-stationary data prior to the consideration of extreme values.

We choose to adopt a Bayesian methodology for our inferential procedures, proceeding via MCMC. The Bayesian framework allows us to produce particularly useful posterior summaries incorporating uncertainty from both the data and the parameters. In particular it enables the calculation of a predictive distribution, which provides a single useful summary of the likelihood of future extremes under the two stated sources of uncertainty (Coles and Tawn, 1996).

Moving from the usual three parameter extreme value models to four parameter models including a Box-Cox parameter necessitates a reparameterization. The theory we exploit to derive our reparameterization is presented in Section 2, including a discussion on the rate of convergence. In Section 3 we outline our reparameterizations and discuss associated inference methods. In Section 4, we illustrate the methodology on simulated data and the aforementioned significant wave height data. A discussion of the work and outstanding issues is given in Section 5.

2 Theory

Suppose $X_1, \ldots, X_n$ are independent and identically distributed according to a probability law with distribution function F_X, twice differentiable with density f_X. Let Y denote these random variables after the application of a Box-Cox transformation; that is $Y = \{X^\lambda - 1\}/\lambda$, $\lambda \in \mathbb{R}$, with distribution function F_Y and density f_Y. Define $M_{X,n} = \max\{X_1, \ldots, X_n\}$. The extremal types theorem (Fisher and Tippett, 1928) states that if there exist sequences of constants $\{a_{X,n} > 0\}$, $\{b_{X,n}\}$ such that as $n \to \infty$

$$\mathsf{P}\left(\frac{M_{X,n} - b_{X,n}}{a_{X,n}} \leq x\right) \xrightarrow{w} G(x)$$

for some non-degenerate limit distribution $G(x)$, then G is necessarily of generalized extreme value type. The symbol '$\xrightarrow{w}$' denotes weak convergence of the distribution functions. The usual premise of extreme value modeling is to assume that this limiting form holds exactly for some finite n.

Let $\{a_{X,n}\}$, $\{b_{X,n}\}$ henceforth specifically denote the normalizing sequences which lead to a GEV$(0, 1, \xi_X)$ limit distribution for the $M_{X,n}$. Because inference on extreme value distributions is performed using un-normalized maxima, the scale and location parameters are estimated approximately as $\sigma_X \approx a_{X,n}$, $\mu_X \approx b_{X,n}$, where n is the block size for the data in question. Smith (1987) shows that the sequences $\{a_{X,n}\}$, $\{b_{X,n}\}$, and the shape parameter ξ_X can be found as follows. Let $h_X(x) = \{1 - F_X(x)\}/f_X(x)$ denote the reciprocal hazard function of the parent distribution F_X. Then

$$b_{X,n} = F_X^{-1}(1 - 1/n) \qquad a_{X,n} = h_X(b_{X,n}) \qquad \xi_X = \lim_{x \to x^F} h'_X(x) \tag{2.1}$$

with $x^F = \sup\{x : F_X(x) < 1\}$, i.e. the upper end point of the distribution. It follows from (2.1) that ξ_X is also given by $\xi_X = \lim_{n\to\infty} h'_X(b_{X,n})$. In the finite sample, $\xi_{X,n} = h'_X(b_{X,n})$ is the so-called penultimate approximation (Fisher and Tippett, 1928; Smith, 1987) to the shape parameter, which is approximately the value of this parameter that is estimated when fitting a GEV distribution to $M_{X,n}$.

The proposal of this paper is to generalize the assumption that $M_{X,n} \dot{\sim} \text{GEV}(\mu_X, \sigma_X, \xi_X)$ to

$$M_{Y,n} = \frac{M_{X,n}^{\lambda} - 1}{\lambda} \dot{\sim} \text{GEV}(\mu_Y, \sigma_Y, \xi_Y),$$

thereby incorporating a form of parametric scale uncertainty into the inference procedure. This gives a four parameter extreme value model, with canonical parameterization $\{\mu_Y, \sigma_Y, \xi_Y, \lambda\}$. The complex nature of the relationships between these parameters however makes direct inference practicably infeasible (see Figure 2 in Section 4 for an illustration). Thus a reparameterization to obtain more orthogonal relationships is necessary. Our strategy for orthogonalization relies upon obtaining $\{a_{Y,n}\}, \{b_{Y,n}\}, \xi_{Y,n}$ in terms of the associated quantities for the original X variables.

Theorem 1. *Given $M_{Y,n} = \{M_{X,n}^{\lambda} - 1\}/\lambda$, and sequences of normalizing constants $\{a_{X,n}\}, \{b_{X,n}\}$ such that*

$$\mathsf{P}\left(\frac{M_{X,n} - b_{X,n}}{a_{X,n}} \leq x\right) \xrightarrow{w} G_X(x) = \exp\left\{-[1 + \xi_X x]_+^{-1/\xi_X}\right\},$$

the sequences $\{a_{Y,n}\}, \{b_{Y,n}\}$ such that

$$\mathsf{P}\left(\frac{M_{Y,n} - b_{Y,n}}{a_{Y,n}} \leq y\right) \xrightarrow{w} G_Y(y) = \exp\left\{-[1 + \xi_Y y]_+^{-1/\xi_Y}\right\}$$

are given by

$$b_{Y,n} = \frac{(b_{X,n})^{\lambda} - 1}{\lambda} \qquad a_{Y,n} = a_{X,n}(b_{X,n})^{\lambda - 1} \tag{2.2}$$

The limiting shape parameter ξ_Y takes the form

$$\xi_Y = \xi_X + \lim_{x \to x^F} \frac{h_X(x)}{x}(\lambda - 1), \tag{2.3}$$

the penultimate approximation to this being given by

$$\xi_{Y,n} = \xi_{X,n} + \frac{a_{X,n}}{b_{X,n}}(\lambda - 1). \tag{2.4}$$

For any differentiable distribution which has $\xi_X \leq 0$ the limiting extreme value type is asymptotically unchanged; that is $\xi_Y = \xi_X$.

See the appendix for a proof. Equations (2.2) and (2.3) are used in Section 3 to motivate our reparameterizations. Note that in the limit under this transformation, when F_X is in the domain of attraction of a Negative Weibull or Gumbel limit, then F_Y remains in this domain of attraction; only those distributions which have a Fréchet limit can be coerced into a different domain. However, as we are never practically in the limit, and $h_X(x)/x > 0$ for $x > 0$, values of λ other than 1 will alter the penultimate approximation and thus change our practical estimation of the shape parameter for the transformed variables regardless of domain of attraction.

It was noted in Section 1 that the rate of convergence to the limiting extreme value distribution may be altered by a power transformation. For $M_{X,n}$ this rate is $\max\{O(|h'_X(b_{X,n}) - \xi_X|), O(n^{-1})\}$, see Smith (1987). An improved rate of convergence will therefore be achieved if $O(|h'_Y(b_{Y,n}) - \xi_Y|) < O(|h'_X(b_{X,n}) - \xi_X|)$, subject to $O(|h'_X(b_{X,n}) - \xi_X|) > O(n^{-1})$. By (2.1), (2.3) and (2.4),

$$|h'_Y(b_{Y,n}) - \xi_Y| = \left| h'_X(b_{X,n}) + \frac{h_X(b_{X,n})}{b_{X,n}}(\lambda - 1) - \lim_{x \to x^F} \left\{ h'_X(x) + \frac{h_X(x)}{x}(\lambda - 1) \right\} \right| \tag{2.5}$$

$$= |h'_X(b_{X,n}) - \xi_X| \left| 1 + (\lambda - 1) \frac{\frac{h_X(b_{X,n})}{b_{X,n}} - \lim_{x \to x^F} \frac{h_X(x)}{x}}{h'_X(b_{X,n}) - \xi_X} \right|. \tag{2.6}$$

Equation (2.6) demonstrates accelerated convergence under the transformation if the second term on the RHS improves the order. This is the case for any value of λ which gives convergence of this second term to 0. In particular this means there is a sequence of λ values, denoted $\{\lambda_n^*\}$ and given by

$$\lambda_n^* \sim 1 - \frac{h'_X(b_{X,n}) - \xi_X}{\frac{h_X(b_{X,n})}{b_{X,n}} - \lim_{x \to x^F} \frac{h_X(x)}{x}}, \tag{2.7}$$

which provide the best rate of convergence under any such transformation. Below we provide illustrations for four different classes of distribution, largely following the examples laid out in Smith (1987). We make the corresponding assumptions that the relationships in Examples 1–3 are twice-differentiable, in the sense that we can differentiate termwise without affecting the O-term representation. Table 1 summarizes the shape parameters for these examples, alongside the order of convergence of the penultimate approximations. Also detailed are values of λ, denoted λ^*, which provide an improved rate of convergence. Note that these values are the limiting values of the sequence $\{\lambda_n^*\}$, where such a limit exists.

Example 1. $x^F = +\infty;\ \alpha, \beta, \epsilon, C > 0;\ D \in \mathbb{R}$

$$1 - F_X(x) = Cx^{-\alpha} \left\{ 1 + Dx^{-\beta} + O(x^{-\beta-\epsilon}) \right\}$$

This class belongs to the Fréchet domain of attraction. Examples include the Pareto, t, F and Cauchy distributions. If $D \neq 0$ then taking $\lambda^* = \beta$ forces the leading term in $|\xi_{Y,n} - \xi_Y|$ to vanish, thus improving the convergence rate.

Example 2. $x^F < +\infty;\ \alpha, \beta, \epsilon, C > 0;\ D \in \mathbb{R}$

$$1 - F_X(x) = C(x^F - x)^{\alpha} \left\{ 1 + D(x^F - x)^{\beta} + O((x^F - x)^{\beta+\epsilon}) \right\}$$

This class belongs to the Negative Weibull domain of attraction. Examples are distributions with bounded upper tails, such as the beta, along with various truncated distributions. Depending on the value of β, the best rate of convergence is either given by $\lambda^* = 1$ ($\beta > 1$), or if $\beta < 1$ the value of λ is asymptotically inconsequential, and in this case the sequence $\{\lambda_n^*\}$ has no limit.

Example 3. $x^F = +\infty;\ \alpha > -1;\ \epsilon > 0;\ C > 0$

$$h_X(x) = \frac{1 - F_X(x)}{f_X(x)} = Cx^{-\alpha}\{1 + O(x^{-\epsilon})\}$$

This class belongs to the Gumbel domain of attraction. Examples include exponential ($\alpha = 0$), normal ($\alpha = 1$), lognormal ($\alpha = 0$), Weibull ($\alpha = \gamma - 1$, for Weibull shape parameter γ), and gamma ($\alpha = 0$). Taking $\lambda^* = \alpha + 1$ improves the rate of convergence, again via elimination of the leading order term in $|\xi_{Y,n} - \xi_Y|$.

In particular note that for the normal distribution $\lambda^* = 2$ leads to faster convergence, the rate being improved from $O((\log n)^{-1})$ to $O((\log n)^{-2})$. More generally for sub-asymptotic levels, when (2.7) is used to obtain the appropriate sequence, $\lambda_n^* \nearrow 2$ as $n \to \infty$. This example is revisited in Section 4.1.2.

Example 4. $x^F = +\infty$ if $\gamma \geq 0$, otherwise $x^F = e^{u - \beta/\gamma}$; $\beta > 0$; $\gamma, u \in \mathbb{R}$

$$1 - F_X(x) = \left[1 + \frac{\gamma}{\beta}(\log x - u)\right]_+^{-1/\gamma}$$

This is the class of 'super-heavy-tailed' log-Pareto distributions, studied in Cormann and Reiss (2009). For $\gamma > 0$ the distribution falls into the domain of attraction of an extreme value distribution if and only if the Box-Cox parameter $\lambda = 0$. This provides the most well-known example of a distribution function outside any domain of attraction: $1 - F_X(x) = 1/\log(x), x > e$, when $\gamma = \beta = u = 1$.

For this class $\lim_{x \to x^F} h'_X(x)$ does not exist if $\gamma > 0$. In such cases (2.6) and (2.7) lack meaning, and one may revert to (2.5) to investigate whether any value of λ which forces the existence of $\lim_{y \to y^F} h'_Y(y)$ can be found. Direct consideration of $|h'_Y(y)|$ in this case, writing x in place of $b_{X,n}$ yields

$$\left|\beta + \gamma(\log x - u) + \gamma - (\lambda - 1) \lim_{x \to x^F} \{\beta + \gamma(\log x - u)\}\right|,$$

the limit of which exists if and only if $\lambda = 0$.

Example	ξ_X	$\xi_{X,n} - \xi_X$	ξ_Y	$\xi_{Y,n} - \xi_Y$	λ^*
1	$1/\alpha$	$\sim \frac{D\beta(\beta-1)}{\alpha^2}(nC)^{-\beta/\alpha}$	λ/α	$\sim \frac{D\beta(\beta-\lambda)}{\alpha^2}(nC)^{-\beta/\alpha}$	β
2	$-1/\alpha$	$\sim \frac{D\beta(\beta+1)}{\alpha^2}(nC)^{-\beta/\alpha}$	$-1/\alpha$	$\sim \frac{D\beta(\beta+1)}{\alpha^2}(nC)^{-\beta/\alpha} - \frac{\lambda-1}{x^F\alpha}(nC)^{-1/\alpha}$	1
3	0	$\sim -\alpha C b_{X,n}^{-\alpha-1}$	0	$\sim C(\lambda - (\alpha+1))b_{X,n}^{-\alpha-1}$	$\alpha + 1$
4	$\beta^\dagger\ \gamma^\diamond$	$0^\dagger\ \beta(n^\gamma)^\diamond$	$\lambda\beta^\dagger,\ \gamma^\ddagger\ \gamma^\diamond$	$0^\dagger,\ 0^\ddagger\ \lambda\beta(n^\gamma)^\diamond$	$0^\star$

Table 1: Shape parameters and the leading order terms from the penultimate approximations for Examples 1–4. $^\dagger \Leftrightarrow \gamma = 0$. $^\ddagger \Leftrightarrow \lambda = 0$ for $\gamma > 0$. $^\diamond \Leftrightarrow \gamma < 0$. * for $\gamma \neq 0$, else same rate achieved $\forall\lambda$.

3 Methodology

3.1 Models

To avoid confusion between features of the limiting distributions and the parameters of our estimated model, we re-label the parameters upon which we perform inference. Specifically our modeling set-up for block maxima is

$$M_{X,n} \dot{\sim} \text{GEV}(\beta_X, \alpha_X, \gamma_X) \qquad\qquad M_{Y,n} = \frac{M_{X,n}^\lambda - 1}{\lambda} \dot{\sim} \text{GEV}(\beta_Y, \alpha_Y, \gamma_Y).$$

This provides parameter sets $\boldsymbol{\theta_X} = \{\beta_X, \alpha_X, \gamma_X\}$ and $\boldsymbol{\theta_Y} = \{\beta_Y, \alpha_Y, \gamma_Y\ \lambda\}$. In particular the shape parameter γ_Y is our finite sample approximation to $\xi_{Y,n}$, the penultimate approximation to the limiting shape

parameter ξ_Y. Estimation of the parameter set $\boldsymbol{\theta_Y}$ directly is unwieldy – see Figure 2 in Section 4.1. Our approach to reducing the dependence amongst the parameter set is described in the following section.

The above description pertains specifically to the GEV model, however a common alternative to the block maxima approach in extreme value analysis is to model all data which exceed some high threshold. The two modeling strategies employed for this purpose are (i) model exceedances via the generalized Pareto distribution (Davison and Smith, 1990), or (ii) model exceedances using a non-homogeneous Poisson process (Pickands, 1971). Case (i) is essentially a reformulation of case (ii), so we discuss here only the latter approach. The formal asymptotic justification for the Poisson process model is that if we have a sequence of two-dimensional point processes

$$P_n = \left\{ \left(\frac{X_i - b_{X,n}}{a_{X,n}}, \frac{i}{n+1} \right) : i = 1, \ldots, n \right\},$$

then on $(x_F^*, \infty) \times (0, 1)$, where $x_F^* = \lim_{n\to\infty}\{x_F - b_{X,n}\}/a_{X,n}$ with $x_F = \inf\{x : F_X(x) > 0\}$, $P_n \to P$, a Poisson process with intensity measure

$$\Lambda((x, \infty) \times (a, b)) = (b - a) \left[1 + \xi_X x\right]_+^{-1/\xi_X} .$$

The normalizing constants $\{a_{X,n}\}, \{b_{X,n}\}$ and the shape parameter ξ_X are exactly as before, thus for statistical inference on un-normalized data we model using a three parameter non-homogeneous Poisson process, denoted $\mathrm{PP}(\beta_X, \alpha_X, \gamma_X)$, with intensity measure

$$\Lambda((x, \infty) \times (c, d)) = (d - c) \left[1 + \frac{\gamma_X}{\alpha_X}(x - \beta_X)\right]_+^{-1/\gamma_X} \qquad 0 \le c < d \le 1.$$

This parameterization is easily unified with that of the GEV. Both GEV and point process methods are considered in our examples in Section 4. In what follows, reference to a '3 parameter model' relates directly to traditional extreme value models whose likelihoods are given by Equations (3.4) and (3.5). Reference to a '4 parameter model' pertains to our extension.

3.2 Reparameterization

When fitting a $\mathrm{GEV}(\beta_Y, \alpha_Y, \gamma_Y)$ distribution to $M_{Y,n}$ the parameters (β_Y, α_Y) will be estimating $(b_{Y,n}, a_{Y,n})$. This is a direct consequence of $a_{Y,n}, b_{Y,n}$ being specifically the sequences which give a $\mathrm{GEV}(0, 1, \xi_Y)$ limit distribution. Therefore Theorem 1 leads to the reparameterizations

$$\beta_Y = \frac{\beta_X^\lambda - 1}{\lambda} \qquad\qquad \log \alpha_Y = (\lambda - 1) \log \beta_X + \log \alpha_X. \tag{3.1}$$

For γ_Y the situation is slightly more subtle. Equation (2.4) suggests taking

$$\gamma_Y = \gamma_X - \frac{\alpha_X}{\beta_X} (\lambda - 1) . \tag{3.2}$$

However Smith (1987) shows via a mean value theorem argument that the estimable value of the shape parameter is $h'_Y(y_0)$ for $y_0 \in [\min\{b_{Y,n}, b_{Y,n} + sh'_Y(b_{Y,n})\}, \max\{b_{Y,n}, b_{Y,n} + sh'_Y(b_{Y,n})\}]$, and s a value in the support of the extreme value distribution. Thus the parametric form (3.2) which is motivated by (2.4) is not strictly appropriate, and the discrepancy between $b_{Y,n}$ and y_0 can be sufficiently large that the structure (3.2) is a poor choice. We do not know where the value y_0 lies. This presents a problem finding a satisfactory theoretical solution to the ratio in (3.2) which multiplies $\lambda - 1$.

To overcome this we have adopted the pragmatic solution of setting

$$\gamma_Y = \gamma_X - c(\lambda - 1), \tag{3.3}$$

where c is a fixed value estimated prior to inference. The value represents the slope in the pairwise profile likelihood for $\{\gamma_Y, \lambda\}$, therefore we estimate it via calculating the profile (log-)likelihood, $\mathrm{P}\ell(\gamma_Y, \lambda)$ on a

fine grid and performing a weighted least squares fit to the grid points in order to extract this slope. The weights are chosen at grid point i to be $\exp[-2\{\mathrm{P}\ell(\hat{\gamma}_Y, \hat{\lambda}) - \mathrm{P}\ell(\gamma_Y, \lambda)_i\}]$, thus ensuring that the ridge of high likelihood dominates the fit and reduces sensitivity of the resulting estimate to the choice of grid. Note that the calculation of $\mathrm{P}\ell(\gamma_Y, \lambda)$ over a particular region of interest presents no difficulties, but full inference from the likelihoods for $\boldsymbol{\theta_Y}$ is infeasible. This two-step approach to the reparameterization has proven to work well in practice.

3.3 Inference

The likelihood functions for a general $\mathrm{GEV}(\beta, \alpha, \gamma)$ distribution, and $\mathrm{PP}(\beta, \alpha, \gamma)$ above a threshold u are given for m data points by

$$L_{\mathrm{GEV}}(\beta, \alpha, \gamma) = \prod_{i=1}^{m} \exp\left\{ -\left[1 + \frac{\gamma}{\alpha}(x_i - \beta)\right]_+^{-1/\gamma} \right\} \frac{1}{\alpha}\left[1 + \frac{\gamma}{\alpha}(x_i - \beta)\right]_+^{-1/\gamma - 1} \tag{3.4}$$

$$L_{\mathrm{PP}}(\beta, \alpha, \gamma) = \exp\left\{ -N_{\mathrm{B}}\left[1 + \frac{\gamma}{\alpha}(u - \beta)\right]_+^{-1/\gamma} \right\} \prod_{i=1}^{m} \frac{1}{\alpha}\left[1 + \frac{\gamma}{\alpha}(x_i - \beta)\right]_+^{-1/\gamma - 1}, \tag{3.5}$$

respectively. In (3.5), the value N_{B} allows manipulation of the parameterization of the Poisson process. A particularly convenient choice when we have a natural 'block length', n, in mind for the data (such as 365 for daily data if interest lies in return level summaries based on annual maxima) is to select N_{B} such that $n = m/N_{\mathrm{B}}$. This produces a Poisson process parameterization which is identical to that of the GEV for M_n. To extend these likelihoods to the 4 parameter case simply requires that u, x_i are replaced by $\{u^\lambda - 1\}/\lambda, \{x_i^\lambda - 1\}/\lambda$, and that each term in the product is multiplied by the Jacobian $x_i^{\lambda - 1}$.

Equations (3.1) and (3.3) represent our reparameterizations of $\{\beta_Y, \log \alpha_Y, \gamma_Y\}$ in terms of a new set of parameters $\{\beta_X, \log \alpha_X, \gamma_X, \lambda\}$. As the first three link clearly to inference for M_X, this allows selection of good choices for parameter starting values by commencing initially with a 3 parameter fit. In our algorithms vague Gaussian priors and Gaussian random walk sampling are used for $\beta_X, \log \alpha_X, \gamma_X$, and a uniform prior with independent sampling for λ. The parameter range for λ is informed by inspection of the profile likelihood $\mathrm{P}\ell(\gamma_Y, \lambda)$.

The algorithm includes the constraint that if $\lambda < 0$, $\gamma_Y < 0$, since the former implies a finite upper end point to the distribution, which is only the case when the latter also holds. Furthermore in the case $\lambda < 0$ this upper end point is $\{(x^F)^\lambda - 1\}/\lambda \leq -1/\lambda$, thus we also impose the constraint that the upper end point of the fitted GEV is $\beta_Y - \alpha_Y/\gamma_Y \leq -1/\lambda$.

It was found that setting $N_{\mathrm{B}} \approx m$, the number of threshold exceedances, in (3.5) improved the mixing properties of the chain. This presents no major difficulties since in general the parameters of a Poisson process corresponding to M blocks of data $\{\beta_M, \alpha_M, \gamma_M\}$ are linked to those of N blocks of data $\{\beta_N, \alpha_N, \gamma_N\}$ via

$$\gamma_M = \gamma_N = \gamma \qquad \beta_N = \beta_M - \frac{\alpha_M}{\gamma}\left(1 - \left(\frac{M}{N}\right)^\gamma\right) \qquad \alpha_N = \alpha_M \left(\frac{M}{N}\right)^\gamma.$$

A reason for improved mixing of GEV parameters under this adjustment is that if there are more data points than the number of blocks then not all the data are providing information on the block maxima parameters; if there are fewer data points than the number of blocks, then this suggests we have incomplete information for these parameters, i.e. the sample of block maxima is censored below.

The output of the MCMC leads to inference on return levels through posterior distributions on specific quantiles, and via the predictive distribution. The $1/p$ block return level, $x_{1/p}$, which is the $1 - p$ quantile of the distribution is found via

$$x_{1/p} = \lambda\left[y_{1/p} + 1\right]^{1/\lambda}, \tag{3.6}$$

where

$$y_{1/p} = \beta_Y - \frac{\alpha_Y}{\gamma_Y}\left[1 - \{-\log(1-p)\}^{-\gamma_Y}\right].$$

The $1/p$ block predictive return level, denoted $x^*_{1/p}$, which corresponds to the $1-p$ quantile of the predictive distribution for $M_{X,n}$, is found by numerically solving

$$\mathsf{P}(M_{X,n} \leq x^*_{1/p}) = \int \mathsf{P}(M_{Y,n} \leq \{{x^*_{1/p}}^{\lambda} - 1\}/\lambda | \boldsymbol{\theta_Y}) p(\boldsymbol{\theta_Y} | \boldsymbol{M_{X,n}})\, d\boldsymbol{\theta_Y} = 1 - p.$$

In practice, this is approximated through a discrete integral over the MCMC output for $\boldsymbol{\theta_Y}$.

4 Examples

4.1 Simulated Data Examples

Two examples are presented. The first illustrates behavior when an exact extreme value distribution is recoverable through a power transformation. The second presents the case of the normal distribution, demonstrating the practical effect of the differing rates of convergence for transformed and untransformed variables. For each example the burn-in period was 1000 iterations, with our reported analyses based on the subsequent 10000 draws.

4.1.1 Pre-Transformed Extreme Value Model

Data were simulated from a NH Poisson process with parameters $\{\beta, \alpha, \gamma\} = \{15, 1.5, -0.25\}$ and the threshold u was fixed by the parameters so that $\Lambda((u, \infty) \times (0, 1)) = 100,000$. The data were generated on the basis of 1000 blocks, i.e. taking N_{B} in (3.5) to be 1000. Three sub-samples of these data were analyzed:

1. Block maxima: 1000 maxima taken of blocks of length 100. These data are exactly GEV$(15, 1.5, -0.25)$ distributed.

2. Largest 1000 data: threshold selected to retain only the largest 1000 points. Owing to the threshold stability property of the Poisson process, these still have a PP$(15, 1.5, -0.25)$ distribution.

3. All data exceeding the smallest block maximum: threshold selected to be equal to the minimum data point in dataset 1. This gave 6847 data points. Again these are PP$(15, 1.5, -0.25)$ distributed.

As a testing ground for the ability of the methodology to detect a 'true' value of λ when one exists, a square transformation was pre-applied to datasets 1, 2 and 3, thus they no longer followed the exact extreme value distributions from which they were generated; these distributions being recoverable, up to location and scale shifts, by taking $\lambda = 0.5$.

Figure 1 displays the posterior distributions for λ in each of the three scenarios. The ranges of the uniform priors for λ are detailed in the caption. Modes around $\lambda = 0.5$ are detectable in (a) and (c) (datasets 1 and 3) with the latter being much the more concentrated density. The least information on λ is obtained from dataset 2. This is explained by the relative extremity of the data. The more extreme the data, the more the standard asymptotic convergence arguments apply. That is, with dataset 2 in particular, the process is approximately Poisson regardless of the transformation since we are still considering the largest 1% of a sample which is in the domain of attraction of an extreme value distribution. Dataset 3 contains a larger amount of data, with the additional data being less extreme than that of dataset 2, thus producing the most informative posterior.

Figure 2 displays the pairwise empirical posteriors from the MCMC output. The first two rows exhibit pairs from the new parameters $\{\beta_X, \log(\alpha_X), \gamma_X, \lambda\}$, whilst the bottom two rows present the implied posteriors for the original parameter set $\{\beta_Y, \alpha_Y, \gamma_Y, \lambda\}$. It is clear from these figures that no meaningful inference could be performed without the reparameterization.

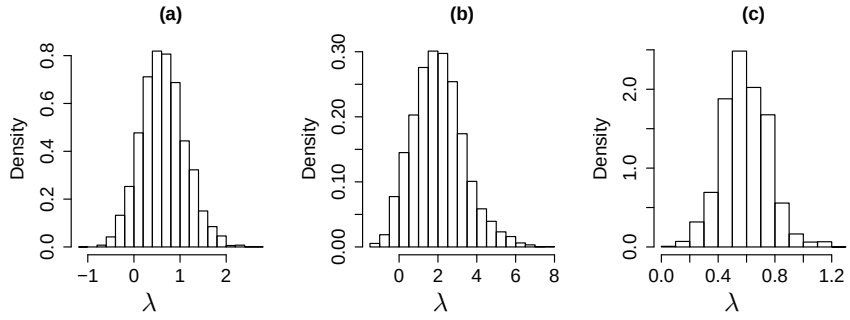

Figure 1: Transformed Extreme Value model example: Posteriors for λ from (a) dataset 1, prior range $[-2, 3]$; (b) dataset 2, prior range $[-2, 8]$; (c) dataset 3, prior range $[0, 2]$.

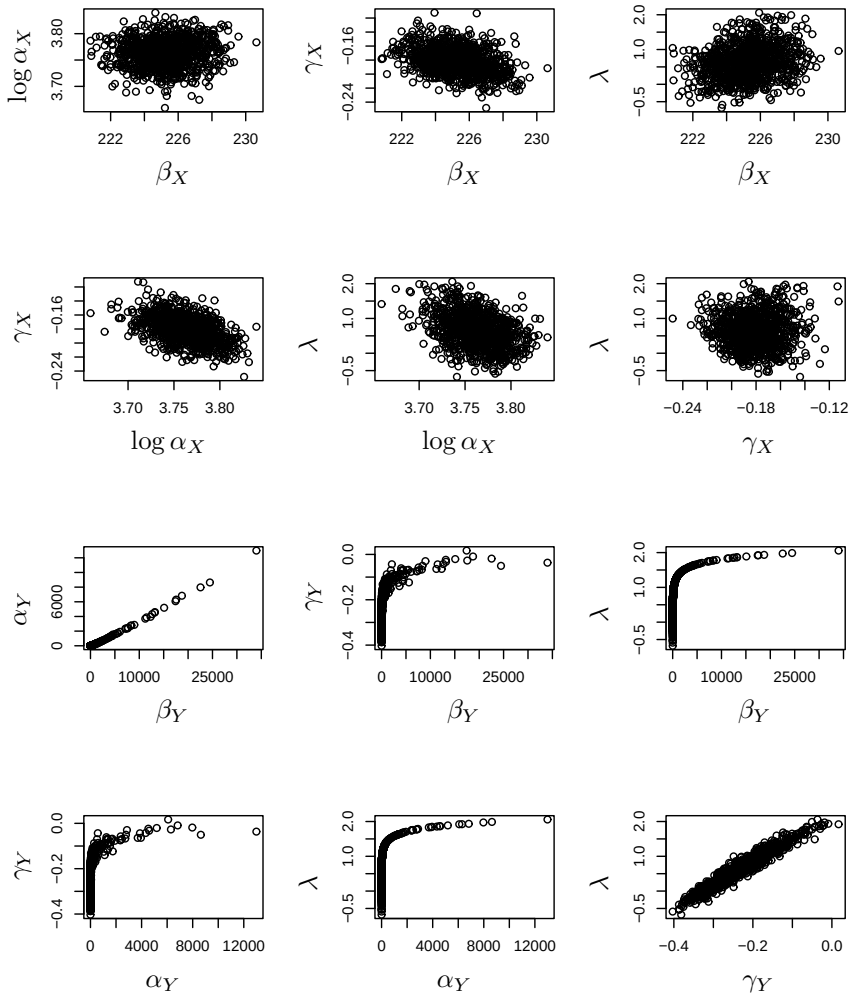

Figure 2: Transformed Extreme Value model example: Pairwise empirical posteriors for the new parameters $\{\beta_X, \log \alpha_X, \gamma_X, \lambda\}$ (top two rows), and the implied posteriors for the original parameters $\{\beta_Y, \alpha_Y, \gamma_Y, \lambda\}$ (bottom two rows).

4.1.2 Normal Distribution

The data simulated were 100,000 truncated (at 0) $\mathrm{N}(0,1)$ variables, i.e. such that $F_X(x) = 2\Phi(x) - 1, x > 0$. As in the example of Section 4.1.1 three datasets were obtained from these:

1. Block maxima: 1000 block maxima taken over block length 100.
2. 1000 largest data points.
3. All data points above the smallest block maximum. There were 8066 such points.

Figure 3 presents the posteriors for λ in each case. The pattern of information contained on λ from each dataset is similar to the previous example, for the reasons formerly described. In Figure 3 (a) there is a mode just below $\lambda = 2$, in (c) the peak is around $\lambda = 1.5$. These values fit well with the theory. The location normalizing constant for the truncated normal distribution is $b_{X,n} \sim (2\log n)^{1/2} - (1/2) \times (2\log n)^{-1/2}(\log \pi) + \log\log n) \approx 2.6$ when $n = 100$. At this sub-asymptotic level, the value of λ_n^* from (2.7), using the first four leading terms in h_X/x and h_X' is 1.86. For the third dataset we are at an even lower asymptotic level. Here, replacing $b_{X,n}$ in the calculation with the threshold, 1.75, gives $\lambda_n^* = 1.48$. Both of these agree with the evidence in the posterior for λ.

Figure 4 displays the return level summaries derived from the analysis, including the true return level curve calculated by solving $F_X(x_{1/p})^{100} = 1 - p$. Posterior return level summaries are displayed pointwise, whilst the predictive distributions are given as curves. In Figures 4 (a) and (c) it can be observed that the 3 parameter model produces biased estimates of the return levels, the true value falling far outside the posterior credible interval. In Figure 4 (b) the true value is just covered by the interval. These results are an indication of the very slow convergence of the Normal distribution to the extreme value limit. From the posteriors for λ there is certainly evidence that accelerated convergence is obtained from the 4 parameter model. The bias in return level estimation compared to the 3 parameter case is reduced, but has not disappeared. The true values of the return level lie within each of the credibility intervals for the 4 parameter models. This is in part down to the faster convergence, although the extra uncertainty involved plays a role as well.

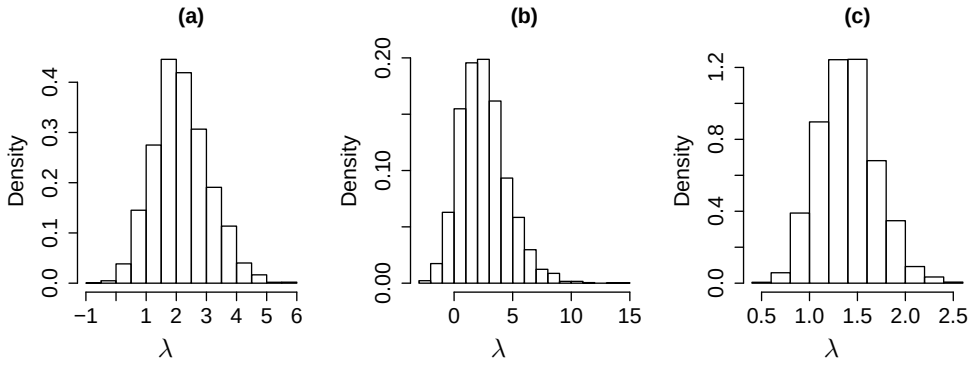

Figure 3: Truncated Normal example: Posteriors for λ, for (a) dataset 1, prior range $[-1, 6]$; (b) dataset 2, prior range $[-3, 15]$; and (c) dataset 3, prior range $[0, 3]$.

4.2 Wave Example

The data are measured significant wave heights (H_s) for an unnamed location in the North Sea. There were just over 33 years of measurements available, with 8 measurements per day recording H_s over continuous 3 hour time periods. Our analysis is restricted to a single season to ensure approximate stationarity, taking the winter period (13 weeks beginning on 1 December each year) as this generally represents the period when almost all extreme events arise.

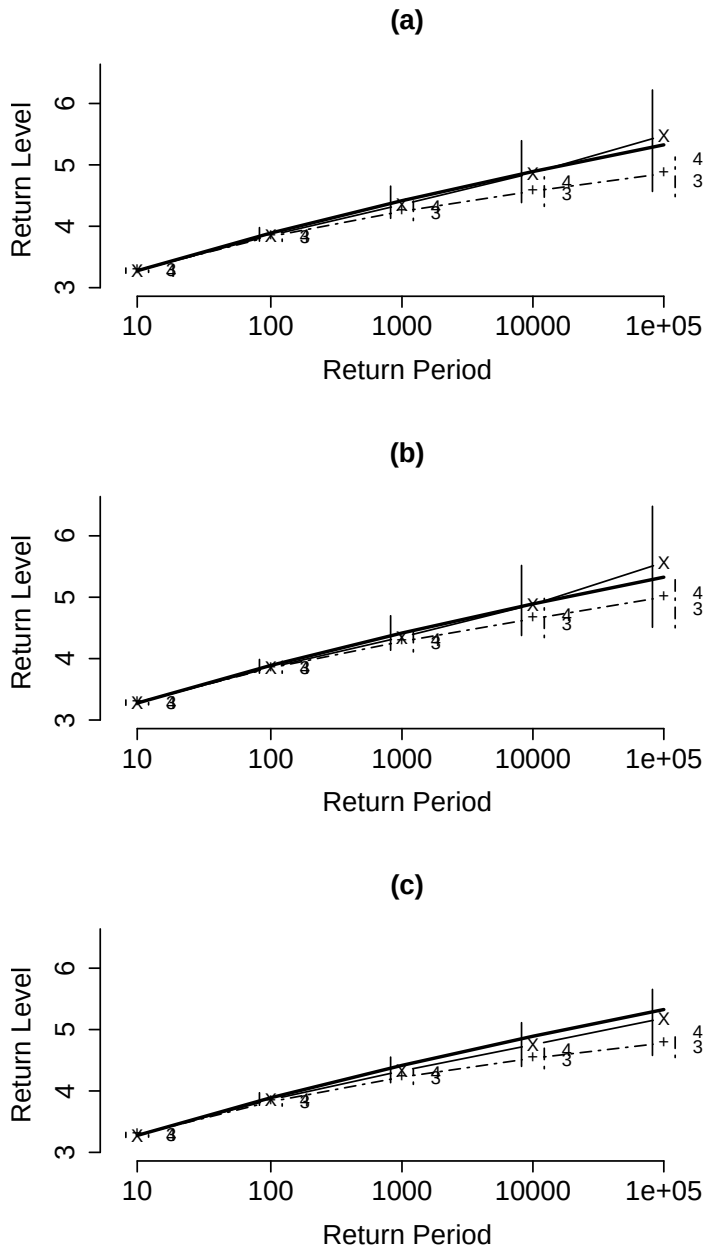

Figure 4: Truncated Normal example: Posterior and predictive return level summaries for (a) dataset 1, (b) dataset 2, (c) dataset 3. The solid bold line represents the true return level based on the truncated normal cdf; '3', '4' denote posterior median return levels of the 3 and 4 parameter models respectively; dashed / solid vertical lines: 3 / 4 parameter model 95% credibility interval; dashed / solid connected lines: 3 / 4 parameter model predictive return levels.

We again examined the data in three ways:

(i) Weekly maxima, corresponding to a block size of $8 \times 7 = 56$ observations. There were 433 data points in total.

(ii) Cluster maxima above an 80% threshold. Runs method declustering (Smith and Weissman, 1994) was used, with a separation of 6 consecutive sub-threshold values deemed to define a new cluster. There were 562 data points.

(iii) Cluster maxima above an 60% threshold, using the same declustering procedure as (ii). There were 618 data points.

In each case both the usual 3 parameter model and the appropriate proposed 4 parameter model (GEV for (i), point process for (ii) and (iii)) were fitted. Our analyses are again based on 10000 MCMC samples following a 1000 iteration burn-in period. Figure 5 displays the profile likelihoods for $\{\gamma_Y, \lambda\}$ and the posterior distributions of λ in each scenario. As with the simulated data, there is more information on λ for less extreme data, evidenced by plots (b) and (f) compared with (d). It is interesting to note that for the 4 parameter GEV model, the slope c in (3.3) was estimated as 0.23, showing how the parameterization (3.3) ties in with the different shape parameters for H_s ($\hat{\gamma}_X = -0.12$) and H_s^2 ($\hat{\gamma}_X = 0.11$) mentioned in Section 1: $0.11 = -0.12 + 0.23 \times (2 - 1)$.

Figure 6 displays QQ plots for each of the fits; here the 'fitted' quantile is defined to be the median of the pointwise quantile posterior distributions, i.e. the median of $x_{1/p}$, for $x_{1/p}$ given by (3.6). Each of the fits appears reasonable, and considering that $\lambda = 1$ is plausible under each of the posteriors, this is perhaps not too surprising. However in each case, there is some evidence that the very upper tail is modeled slightly better by the 4 parameter model.

Posterior summaries of the return levels from analysis (iii) are displayed in Figure 7, where increasing disparity of estimates with lengthening return period can be observed. The corresponding plots for analyses (i) and (ii) have been omitted for clarity, but show similar general trends with greater uncertainty for analysis (ii) and lesser for analysis (i). In particular observe that the medians of the posterior return level distribution for the 100 and 1000 winter return periods under the 4 parameter model lie into the upper tail of the same distributions under the 3 parameter model. From the motivating example in Section 1 it is clear why these discrepancies occur: the H_s data were estimated as light-tailed, with a statistically significant negative shape parameter (taking a 5% significance level); the H_s^2 data were estimated to be heavy-tailed, with a statistically significant positive shape parameter. Such different tail behavior will naturally lead us to different conclusions. The posteriors for λ show that we might reasonably extrapolate on either scale, amongst other possibilites; the 4 parameter model combines all such plausible scenarios to build up what would appear to be a more accurate assessment of the uncertainty associated with these extrapolations.

5 Discussion

The paper has presented a parametric method for incorporating the uncertainty surrounding the scale of extrapolation in extreme value analysis. Reparameterizations which allow inference under the model have been derived, justified by the theory of normalizing constants for the limiting distribution of block maxima. Examples have demonstrated the ability of the methodology to detect the 'true' value of λ where one exists, for the case of finite block size / threshold. As either of these quantities tend to infinity, information on λ decreases, since there is little to be gained from a transformation.

The fact that there may not always be significant information on λ poses the question whether it is always necessary to incorporate this uncertainty. In Theorem 1 we noted that in the case where $\xi_X \leq 0$ with $x^F > 0$, the shape parameters $\xi_{Y,n} \to \xi_X$ as the data become more extreme, since $\lim_{x \to x^F} h_X(x)/x = 0$. In such a case, where all our data are far into the upper tail, the mean squared error of the 4 parameter model is likely to exceed that of the 3 parameter case. An ill-determined posterior for λ may be one indication that

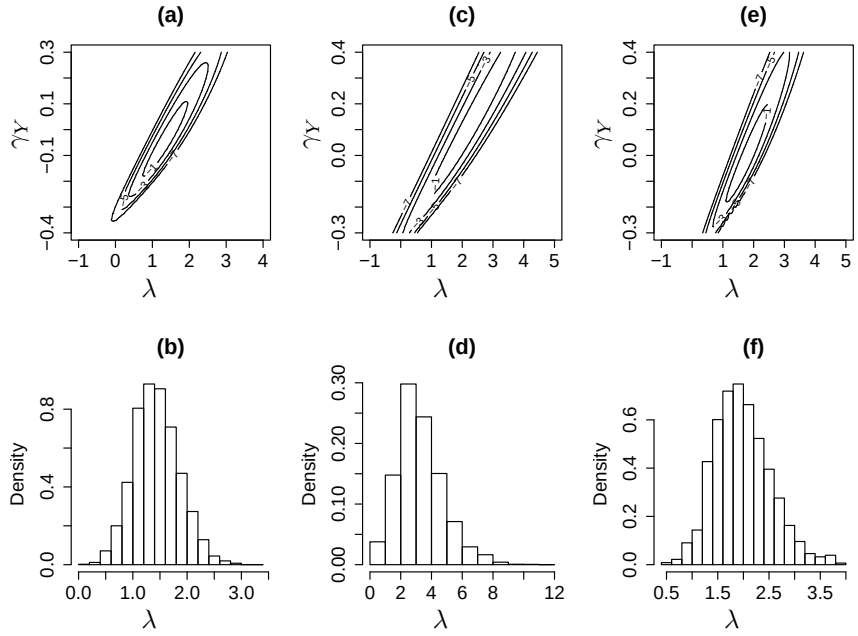

Figure 5: H_s data example: (a), (c), (e) Profile log-likelihoods for $\{\gamma_Y, \lambda\}$, with contours at levels of -1, -3, -5, -7 below the maximum log-likelihood; (b), (d), (f) posteriors for λ for analyses (i), (ii) and (iii) respectively. Prior ranges for λ taken as (i) $[-1, 4]$, (ii) $[-2, 15]$, (iii) [0,5].

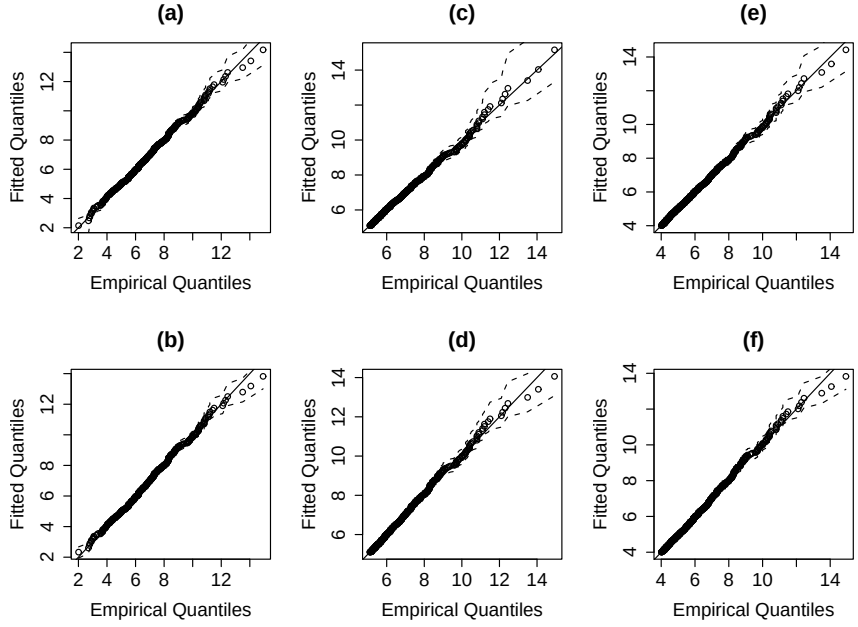

Figure 6: H_s data example: QQ plots for (a), (c), (e) 4 parameter model, (b), (d), (f) 3 parameter model for analyses (i), (ii) and (iii) respectively. Dashed lines represent a 95% pointwise credible interval, formed from the central 95% of the posterior distribution for each quantile.

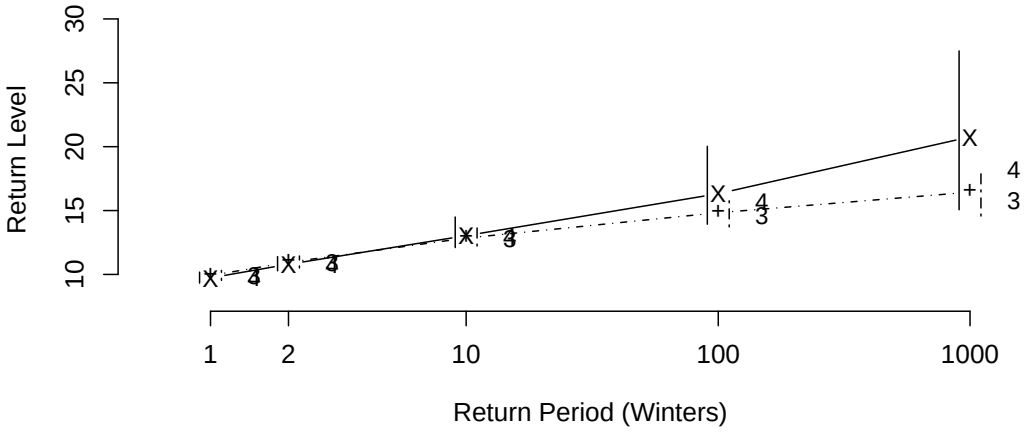

Figure 7: H_s data example: Posterior and predictive return level summaries for H_s, based on both 3 and 4 parameter models for analysis (iii). '3', '4' denote posterior median return levels of the 3 and 4 parameter models respectively; dashed / solid vertical lines: 3 / 4 parameter model 95% credibility interval; dashed / solid connected lines: 3 / 4 parameter model predictive return levels.

utilization of this method adds an unnecessary degree of uncertainty. As with other issues such as threshold selection, rational judgment of the practising statistician is required.

At the other end of the scale, the fact that suitable values of λ may accelerate convergence offers the potential for incorporating more data through lowering of the threshold or contracting of block length. Although we have not specifically explored this here, examples such as the normal data example given in Section 4 demonstrate how this could be worthwhile. Because QQ plots such as those in Figure 6 are easily obtained under both 3 and 4 parameter models, the modeler should be able to determine if there is value in doing this.

A natural question that arises is whether to consider fixing λ if there is strong evidence for a particular value in the posterior. As outlined in Section 2 there are cases where a specific value will accelerate convergence, thus one could assume that the modal value is a suitable one to take. However, the reason that we have a full posterior distribution is that there is genuinely uncertainty in this value. Arguably therefore we mitigate against our errors by keeping this uncertainty. This seems unsatisfying in a world where we value precision in our estimates, but if uncertainty genuinely exists it should not be masked by the pursuit of false precision.

The Box-Cox class of transformations is suitable only for strictly positive data. In the event that interest lies in a dataset for which this is not the case, a location shift prior to analysis would be necessary. One might also in such a case consider different classes of transformation. Cormann and Reiss (2009) for example consider exponential transforms. From our proof in Appendix A it is simple to derive reparameterizations for any monotonic transformation, thus one could exploit this theory in other contexts.

Acknowledgments

JLW would like to acknowledge the EPSRC and Shell Research for funding through a CASE studentship. The authors further acknowledge discussions with Kevin Ewans of Shell International Exploration and Production, concerning the wave data analyzed in Section 4.

A Appendix: Proof of Theorem 1

Denote the transformation $y(x) = \{x^\lambda - 1\}/\lambda$, and the inverse transformation $x(y) = \{\lambda y + 1\}^{1/\lambda}$. The distribution function F_Y is given by

$$F_Y(y) = \mathsf{P}(Y \leq y) = \mathsf{P}(X \leq \{\lambda y + 1\}^{1/\lambda}) = F_X(\{\lambda y + 1\}^{1/\lambda}) = F_X(x(y)).$$

Therefore solving $F_Y(b_{Y,n}) = 1 - 1/n$ for $b_{Y,n}$ yields

$$\begin{aligned} F_X(\{\lambda b_{Y,n} + 1\}^{1/\lambda}) &= 1 - 1/n \\ \{\lambda b_{Y,n} + 1\}^{1/\lambda} &= F_X^{-1}(1 - 1/n) = b_{X,n} \\ b_{Y,n} &= \frac{b_{X,n}^\lambda - 1}{\lambda}. \end{aligned}$$

Denote the Jacobian of the transformation and inverse transformation by

$$J_X(x) := \left|\frac{dy}{dx}\right| = x^{\lambda-1} \qquad\qquad J_Y(y) := \left|\frac{dx}{dy}\right| = \{\lambda y + 1\}^{1/\lambda - 1}.$$

These are linked by $J_Y(y) = \{J_X(x(y))\}^{-1}$. The reciprocal hazard function h_Y is

$$h_Y(y) = \frac{1 - F_Y(y)}{f_Y(y)} = \frac{1 - F_X(x(y))}{f_X(x(y))J_Y(y)} = \frac{h_X(x(y))}{J_Y(y)},$$

which gives

$$a_{Y,n} = h_Y(b_{Y,n}) = \frac{h_X(\{\lambda b_{Y,n} + 1\}^{1/\lambda})}{\{\lambda b_{Y,n} + 1\}^{1/\lambda - 1}} = \frac{h_X(b_{X,n})}{(b_{X,n})^{1-\lambda}} = a_{X,n}(b_{X,n})^{\lambda-1}.$$

To obtain an expression for the shape parameter we require the derivative of the reciprocal hazard function for Y.

$$\begin{aligned} h_Y'(y) &= \frac{d}{dy}\left\{\frac{h_X(x(y))}{J_Y(y)}\right\} \\ &= \frac{J_Y(y)\frac{d}{dy}h_X(x(y)) - h_X(x(y))\frac{d}{dy}J_Y(y)}{J_Y(y)^2}. \end{aligned}$$

By the chain rule,

$$\frac{d}{dy}h_X(x(y)) = J_Y(y)h_X'(x(y)),$$

and

$$J_Y'(y) = J_Y(y)\frac{d}{dx}\frac{1}{J_X(x(y))} = -\frac{J_X'(x(y))}{J_X(x(y))^3}.$$

Thus,

$$\begin{aligned} h_Y'(y) &= \frac{J_Y(y)^2 h_X'(x(y))}{J_Y(y)^2} - \frac{h_X(x(y))J_Y'(y)}{J_Y(y)^2} \\ &= h_X'(x(y)) + h_X(x(y))\frac{J_X'(x(y))}{J_X(x(y))}. \end{aligned}$$

Substituting in $J_X(x) = x^{\lambda-1}$, $J'_X(x) = (\lambda-1)x^{\lambda-2}$ results in

$$h'_Y(y(x)) = h'_X(x) + \frac{h_X(x)}{x}(\lambda - 1).$$

Substituting in $x = b_{X,n}$ gives (2.4); taking the limit as $x \to x^F$ gives (2.3).

For the final statement, $\xi_X = \lim_{x\to x^F} h'_X(x) \leq 0$ implies that $\lim_{x\to x^F} h_X(x)/x = 0$.

References

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, **26**, 211–252.

Coles, S. G. (2001) *An Introduction to the Statistical Modeling of Extreme Values*. Springer-Verlag, London.

Coles, S. G. and Tawn, J. A. (1996) A Bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **45**, 463–478.

Cormann, U. and Reiss, R.-D. (2009) Generalizing the Pareto to the log-Pareto model and statistical inference. *Extremes*, **12**, 93–105.

Davison, A. C. and Smith, R. L. (1990) Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, **52**, 393–442.

Eastoe, E. F. and Tawn, J. A. (2009) Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **58**, 25–45.

Fisher, R. A. and Tippett, L. H. C. (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, **24**, 180–190.

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

Pickands, J. (1971) The two-dimensional Poisson process and extremal processes. *Journal of Applied Probability*, **8**, 745–756.

Smith, R. L. (1987) Approximations in extreme value theory. University of North Carolina, Department of Statistics, Technical Report No. 205.

Smith, R. L. and Weissman, I. (1994) Estimating the extremal index. *Journal of the Royal Statistical Society. Series B (Methodological)*, **56**, 515–528.

Tromans, P. S. and Vanderschuren, L. (1995) Based design conditions in the north sea: Application of a new method. In *Offshore Technology Conference, Houston OTC-7683*.