Posterior Predictive Checks for the Generalized Pareto Distribution Based on a Dirichlet Process Prior

Wilson Moseki Thupeng[*], Boikanyo Mokgweetsi, Thuto Mothupi

Department of Statistics, University of Botswana, Gaborone, Botswana

Email address:
Thupengw@ub.ac.bw (W. M. Thupeng), Mokgweetsib@mopipi.ub.bw (B. Mokgweetsi), mothupit@mopipi.ub.bw (T. Mothupi)
[*]Corresponding author

Abstract: Extreme value modelling is widely applied in situations where accurate assessment of the behavior of a process at high levels is needed. The inherent scarcity of extreme value data, the natural objective of predicting future extreme values of a process associated with modelling of extremes and the regularity assumptions required by the likelihood and probability weighted moments methods of parameter estimation within the frequentist framework, make it imperative for a practitioner to consider Bayesian methodology when modelling extremes. Within the Bayesian paradigm, the widely used tool for assessing the fitness of a model is by using posterior predictive checks (PPCs). The method involves comparing the posterior predictive distribution of future observations to the historical data. Posterior predictive inference involves the prediction of unobserved variables in light of observed data.. This paper considers posterior predictive checks for assessing model fitness for the generalized Pareto model based on a Dirichlet process prior. The posterior predictive distribution for the Dirichlet process based model is derived. Threshold selection is done by minimizing the negative differential entropy of the Dirichlet distribution. Predictions are drawn from the Bayesian posterior distribution by Markov chain Monte Carlo simulation (Metropolis-Hastings Algorithm). Two graphical measures of discrepancy between the predicted observations and the observed values commonly applied in practical extreme value modelling are considered, the cumulative distribution function and quantile plots. To support these, the Nash-Sutcliffe coefficient of model efficiency, a numerical measure that evaluates the error in the predicted observations relative to the natural variation in the observed values is used. Finite sample performance of the proposed procedure is illustrated through simulated data. The results of the study suggest that posterior predictive checks are reasonable diagnostic tools for assessing the fit of the generalized Pareto distribution. In addition, the posterior predictive quantile plot seems to be more informative than the probability plot. Most interestingly, selecting the threshold by minimizing the negative differential entropy of a Dirichlet process has the added advantage of allowing the analyst to estimate the concentration parameter from the model, rather than specifying it as a measure of his/her belief in the proposed model as a prior guess for the unknown distribution that generated the observations. Lastly, the results of application to real life data show that the distribution of the annual maximal inflows into the Okavango River at Mohembo, Botswana, can be adequately described by the generalized Pareto distribution.

Keywords: Dirichlet Process Prior, Generalized Pareto Distribution, Markov Chain Monte Carlo, Peaks Over Threshold, Posterior Predictive Checks

1. Introduction

Modelling extreme values is widely applied in situations, like civil engineering design, where accurate assessment of the behavior of a process at high levels is required. In the statistics literature, most of the works on extreme value analysis and modelling are based on the likelihood or frequentist approach. However, there are several reasons why a statistician might prefer Bayesian methodology to the frequentist approach when modelling extreme values: (i) extreme value data are inherently scarce, making it natural

for the statistics practitioner to consider including other available prior information in the analysis, (ii) the main objective of an extreme value analysis is to estimate the probability of future events reaching extreme levels and predicting high quantiles and, one approach to prediction is through the predictive distribution, which is naturally Bayesian and, (iii) the Bayesian methodology does not require regularity assumptions required by the maximum likelihood and probability weighted moments methods of estimation. It is for these reasons that the present paper focusses on using posterior predictive checks for assessing model-fit for the generalized Pareto distribution, a family of continuous probability distributions that is popularly used to model the tails of distributions of extreme values. The reader is referred to Coles and Powell [7] for a comprehensive review of and new developments in the utility of a Bayesian approach to the inference of extremes.

In applications, there are two widely adopted approaches to modelling extreme values of a time series of independent observations: Block maxima method, in which data are subdivided into sequences of observations of size n each, for some large n, resulting in a series of block maxima, which are commonly modelled using the generalized extreme value (GEV) distribution, and, the peaks over threshold (POT) approach in which all exceedances over a specified high threshold are modelled. In this paper, our main interests is in predicting high quantiles, and, as a result, we consider the POT approach in which all observations in the series $x_1, x_2, \ldots, x_n$ which exceed a high threshold, say t, are modelled.. The POT methods are based on fitting a stochastic model to either the exceedances or the peaks over a threshold [9]. We discuss the fit of the generalized Pareto distribution (GPD), the most popular model when fitting data using the POT method [9, 23, 35]. The advantage of the POT method in modelling extremes is that more observations are incorporated into the analysis [28].

There are many model assessment diagnostics in the literature, but no overall goodness of fit test exists and the scientist should specify his preference [21]. One of the main Bayesian methods for assessing the fit of parametric models involves embedding the parametric family into a larger family called the extended model [4, 6, 24]**.** This method is more useful if the larger family is infinite-dimensional so that the extended model becomes nonparametric [37]. The Dirichlet process by Ferguson is one of the many infinite-dimensional models used in Bayesian inference for generating random distribution functions [13]. It is widely used in Bayesian inference as a prior on a set of probability distributions of random variables, enabling the statistician to consider certain nonparametric problems from a Bayesian approach.

To assess the fit of the model mentioned above, we consider diagnostics based on predictions generated by the model. Predictions are drawn from the Bayesian posterior predictive distribution by Markov chain Monte Carlo simulation (Metropolis-Hastings Algorithm). Comparing the posterior predictive distribution to the observed values is generally referred to as a posterior predictive check [14, 17,

19, 32, 33]. One of the advantages of this approach is that the posterior predictive distribution reflects both the *parametric uncertainty* (via prior specification) and *sampling uncertainty* (via the sampling distribution of the future observation). If the model fits the data well, then its predictions should resemble the data; large discrepancies between the observed data and the predicted observations indicate poor fit. Posterior predictive checks for assessing statistical models has been gaining currency as one of several existing techniques for Bayesian evaluation of model goodness of fit. For example, Mmino et al. [27] have discussed a general statistical procedure for checking the goodness-of-fit of an admixture model to genomic data based on posterior predictive checks [27]. The authors find that the same model fitted to different genomic studies resulted in highly study-specific results when evaluated using PPCs, illustrating the utility of PPCs for model-based analyses in large genomic studies. Another author investigates the sensitivity of the posterior predictive checks to prior specification for Item response theory models and finds that the PPC procedure hit rates appear to be influenced by prior specification, but only in certain circumstances and the effect of prior in formativeness is tied to the choice of discrepancy measure [1]. Fawcett and Walshaw [13] demonstrate predictive inference for return levels of wind speed and sea-surge extremes and recommend the posterior predictive return level as the most convenient and useful representation of a return level which provides practitioners with a point summary capturing estimation uncertainty [13]. A large scale simulation study reveals the superiority of the predictive return level over the other posterior summaries in many cases of practical interest [14]. Despite the utility of posterior predictive checks for assessment of model-data fit, literature shows that not much work has been done in the area of Bayesian inference for extremes values. The present study seeks to address this gap by considering posterior predictive checks for assessing model fitness for the generalized Pareto model based on a Dirichlet process prior.

In this paper graphical PPCs and a numerical measure of discrepancy between predicted observations and observed values are both considered. Two graphical diagnostics are used: the cumulative distribution function of the predicted distribution (PD) overlaid on the estimated GPD and, a Q-Q plot of the predicted quantiles against the observed values. The numerical measure of discrepancy is given by the Nash-Sutcliffe coefficient of model efficiency, which measures the error in the predicted observations relative to the natural variation in the observed values [29].

The paper is organized as follows. Section 2 presents the generalized Pareto model. Section 3 provides derivations of the Dirichlet process, the negative differential entropy of the Dirichlet distribution, a statement of some important results for the negative differential entropy of the Dirichlet, and a description of the method of threshold selection. Derivation of the posterior predictive distribution and model assessment tools are the subject of section 4. Section 5 presents an application on the modelling of annual maximum water

inflows into the Okavango River at Mohembo, Botswana, during the period 1975-2003. Section 6 contains conclusions on the study.

1.1. The Generalized Pareto Model

Let $x_1, x_2, \ldots, x_n$ be a series of n independent random variables with a common distribution function F. Extreme events are defined by identifying a high threshold, say t, with exceedances $\{x_i : x_i > t\}$. Denote by $x_1, x_2, \ldots, x_k$ the k observations exceeding a specified threshold t in a sample of n extreme observations of a process. Threshold selection remains an active area of research in extreme value theory. De Waal proposed the choice of threshold through minimization of the entropy of the Dirichlet process when applying the POT method [10]. We follow this approach.

To formulate the model, consider the asymptotic behavior of the series $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$. With normalizing sequences $\{a_n\}$ and $\{b_n\}$, the complete class of non-degenerate limiting distributions of $\left(X_{(n)} - a_n\right)\big/b_n$ as $n \to \infty$ is the generalized extreme value (GEV) family. This has the distribution function.

$$G(x) = e^{-[1+\gamma(\frac{x-\mu}{\sigma})]^{-\frac{1}{\gamma}}} \qquad (1)$$

defined on the set $\{x : 1 + \gamma(x-\mu)/\sigma > 0\}$, with $-\infty < \mu < \infty, \sigma > 0, -\infty < \gamma < \infty$

The parameters μ and σ are location and scale parameters, while γ is a shape parameter which determines the weight of the tail of G, and hence F. The current study is based on the generalized Pareto distribution which has an interpretation as a limiting distribution similar to the GEV family [9]. In fact, the GPD belongs in the domain of attraction in which the shape parameter in (1) is positive, i.e. $\gamma > 0$. To implement the POT method, denote by

$$P(x-t \le x | X > t) = \frac{F(t+y) - F(t)}{1 - F(t)}$$

$$i.e \qquad F(y|t) = \frac{F(t+y) - F(t)}{1 - F(t)}$$

$$G\left(x|t, \sigma_t, \gamma\right) = 1 - \left(1 + \frac{\gamma(x-t)}{\sigma_t}\right)^{-\frac{1}{\gamma}}, \ x > t; \ \gamma > 0 \quad (2)$$

The distribution of the excesses, $Y_i = X_i - t$ of X over t, given that t is exceeded. Pickands has shown that the conditional distribution of threshold exceedances follows the generalized Pareto distribution, with distribution function (df) given by

Where σ_t and γ are the GEV parameters of model (1), with $\mu = 0$. If we let $x_0 \le \infty$ denote the upper endpoint of F, and define $F(y|t)$ as above, Pickands showed that the GPD is a good approximation of $F(y|t)$ in the sense that

$$\lim_{t \to x} \sup_{0 < y < x_0 - t} \left| F(y|t) - G(x; \sigma_t, \gamma) \right| = 0, \text{ for some fixed } \gamma \text{ and}$$

σ_t, if and only if F is in the domain of attraction of one of the three extreme value limit laws [31]. Inference based on model (2) is generally superior since it applies to more data [8]. Parameters of the model (2) are estimated via MCMC simulation using the Metropolis-Hastings algorithm.

1.2. The Dirichlet Process

Partition the sample space according to the largest k ordered observations exceeding a threshold t: $t < x_{(1)}, x_{(2)}, \ldots, x_{(k)} < \infty$. Let $F(x|t)$ denote the true unknown df of random variable X and $G(x|t)$ be the proposed df with known parameters. $G(x|t)$ is referred to as the base measure of Dirichlet process (DP). Define the probabilities

$$p_i = P(x_{(i-1)} | t < X < x_{(i)} | t) = F\left(x_{(i)} | t\right) - F\left(x_{(i-1)} | t\right) \qquad \text{for}$$

$i = 1, 2, \ldots, k$ and $p_{k+1} = 1 - F\left(x_{(k)} | t\right)$. Clearly, $P_1, P_2, \ldots, P_{k+1}$ are random variables, and have jointly the k-variate Dirichlet distribution with parameters $v_1, v_2, \ldots, v_{k+1}$. That is, $P_1, P_2, \ldots, P_{k+1} \sim D\left(v_1, v_2, \ldots, v_{k+1}\right)$ with density function given by

$$f\left(p_1, p_2, \ldots, p_{k+1}\right) = \frac{\Gamma(v_o)}{\prod_{i=1}^{k+1} \Gamma(v_i)} \prod_{i=1}^{k+1} p_i^{v_i - 1}, \ p_i \ge 0, \sum_{i=1}^{k+1} p_i = 1, v_i > 0, v_o = \sum_{i=1}^{k+1} v_i \qquad (3)$$

Now, if $P_1, P_2, \ldots, P_{k+1}$ are jointly $D\left(v_1, v_2, \ldots, v_{k+1}\right)$ with parameters defined by $v_i = \beta\left\{G\left(x_{(i)} | t\right) - G\left(x_{(i-1)} | t\right)\right\}, i = 1, 2, \ldots k$ and $v_{k+1} = 1 - G\left(x_{(k)} | t\right)$ under the proposed model, with

$G(x|t) = \mathrm{E}\left(F(x|t)\right)$ as a prior for $F(x|t)$, then, a priori, $F(x|t)$ is said to be described through a Dirichlet process with parameters $G(x|t)$ and β, i.e. $F(x|t) \sim DP\left(G(x|t), \beta\right)$. The parameter β is called the

concentration parameter and measures the analyst's belief in $G(x|t)$ as a prior for $F(x|t)$. It has been shown that in order for the negative differential entropy to reach its lower bound for a given k, β has to be proportional to k, say, $\beta = \alpha(k+1)$ for some real $\alpha > 0$. The choice of α, also called the concentration parameter, for a given k is obtained by minimizing the negative differential entropy of the Dirichlet distribution.

From the above setup, it follows that, if we assume a priori that a certain threshold model, say the generalized Pareto, with df $G(x|t)$ best fits the data, then it follows by Bayes Theorem that, a posteriori, $F(x|t)$ is described through a Dirichlet process with parameters $\beta G(x|t) + (k+1) F_k(x)$ and $\beta + k$. Denote this, for $x > t$, by

$$F(x|t) \sim DP\big(\beta G(x|t) + (k+1) F_k(x), \beta + k\big) \quad (4)$$

where $F_k(x) = \dfrac{1}{k+1} \sum_{i=1}^{k} \partial_{x_i}(x)$ is the empirical distribution

$$f(p_1, p_2, \ldots, p_{k+1}) = \frac{\Gamma(v_o)}{\prod\limits_{i=1}^{k+1} \Gamma(v_i)} \prod_{i=1}^{k+1} p_i^{v_i - 1}, \;\; p_i \geq 0, \sum_{i=1}^{k+1} p_i = 1, v_i > 0, v_o = \sum_{i=1}^{k+1} v_i$$

Then, Honkela showed that the negative differential entropy, defined as $J_{k+1} = E \ln f(P_1, P_2, \ldots, P_{k+1})$, for the Dirichlet distribution is given by

$$J_{k+1} = \ln \Gamma(v_0) - \sum_{i=1}^{k+1} \ln \Gamma(v_i) + \sum_{i=1}^{k+1} (v_i - 1)\{\Psi(v_i) - \Psi(v_0)\} \quad (5)$$

where $v_0 = \sum_{i=1}^{k+1} v_i$ and $\Psi(v) = \dfrac{d}{dv} \ln \Gamma(v)$, the digamma function, and $E\{\ln P_i\} = \Psi(v_i) - \Psi(v_0) \, \forall i$ [20].

1.3.1. Some Results for the Negative Differential Entropy of the Dirichlet Distribution

We have established the following results for the NDE of the Dirichlet distribution. The proofs have been left out for lack of space.
  i. The NDE is a convex function of k.
  ii. The NDE reaches a minimum when $v_i = 1 \, \forall \, i$.

The lower bound on the NDE is given by $J_{k+1}^* = \ln \Gamma(k+1), k = 1, 2, \ldots, n-1$.

For the NDE to reach its lower bound for a given k, β has to be proportional to k, say, $\beta = \alpha(k+1)$ for some real $\alpha > 0$.

The results are obtained by minimizing the NDE in (2.5) with respect to the parameters $v_1, v_2, \ldots, v_{k+1}$ of the k-variate

function for the random variable X and $G(x|t)$ is the distribution function for the proposed model.

1.3. The Negative Differential Entropy of the Dirichlet Distribution

Shannon's seminal discovery of a quantitative measure for entropy came in connection with communication theory, and very soon his concept of information had a pervasive influence on several other disciplines [22]. These disciplines include statistics. According to Shannon, a measure of the entropy (uncertainty) of a probability distribution reflects the amount of ignorance in our state of knowledge. Conversely, a measure of the negative entropy (lack of uncertainty) of a probability distribution reflects the amount of information in our state of knowledge. Though originally defined for a discrete random variable, Shannon's entropy can be extended to the case of a continuous random variable. Then the entropy is referred to as differential entropy, and its negative as *negative differential entropy*. We concentrate on the latter.

Suppose the random variables $P_1, P_2, \ldots, P_{k+1}$ have jointly the k-variate Dirichlet distribution, $D(v_1, v_2, \ldots, v_{k+1})$, with probability density function (pdf)

Dirichlet distribution and solving at zero for $v_i, i = 1, 2, \ldots, k+1$, and using the fact that $\sum_{i=1}^{k+1} v_i = \beta$ from the definition of v_i. The proofs have been left out for lack of space.

1.3.2. Threshold Selection

The results stated in section 3.2 are useful in selecting the threshold when applying the POT method. The threshold is selected by finding the number of observations, k, above the threshold t that minimizes the negative differential entropy of the Dirichlet distribution. In practice, plot the relative NDE against k and select the optimal k as that k which corresponds to the minimum of the relative NDE. The relative NDE is obtained by dividing the NDE in (5) by its lower bound, $J_{k+1}^* = \ln \Gamma(k+1), k = 1, 2, \ldots, n-1$. The parameters $v_1, v_2, \ldots, v_{k+1}$, and (by implication) α or β in (5) are estimated by MCMC simulation.

Remark 1: Mazzuchi applied the Dirichlet process to introduce a Bayesian technique for assessing goodness of fit of models [28]. De Waal and Beirlant showed that the relative negative differential entropy of the Dirichlet process could possibly be used as a tool for model selection [11]. That is, given two competing models, Model 1 and Model 2 with relative NDE $J_{k_1}(Model\,1)$ and $J_{k_2}(Model\,2)$ respectively, the analyst would select the model with the

minimum NDE, $J_{k_i}, i = 1, 2$ (i.e. the model that gives maximum information about the data generating mechanism).

1.3.3. Selection of Concentration Parameter

Once the threshold has been selected, we can determine the concentration parameter α. The choice of α for a given k is obtained by minimizing the negative differential entropy of the Dirichlet distribution. This leads to various selected values of α corresponding to various minimum values of the negative differential entropy for different values of k. Thus, to find the optimal value of α, plot the selected α's against k as in Figure 4.

2. The Predictive Distribution and Model Assessment

2.1. The Predictive Distribution

Remark 2: When $k = 1$, the Dirichlet $D(v_1, v_2)$ distribution is identical to the $beta(v_1, v_2)$ distribution so that Equation (2.3) reduces to the pdf of the beta distribution. Therefore, from the Dirichlet process, we that marginally, $F(x|t) \sim beta(a_k, b_k)$, where

$a_k = \beta G(x|t) + (k+1) F_k(x)$ and $b_k = \beta + k - a_k$ (by (4)).

$H(x_f | t, data) = P(X_f < x_f | t, data)$ From theory, the predictive distribution of a future observation x_f is given by

$$H(x_f | t, data) = \int_0^1 h(u|t, data) P(X_f < x_f | u) du \quad (6)$$

If we let $F(x|t) = u$, then we have that

where $h(u|t, data)$ denotes the posterior density of $F(x|t) = u$ given the data $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$, which is $beta(v_1, v_2)$, and $P(X_f < x_f | u) = u$. This implies that

$$H(x_f | t, data) = \frac{\Gamma(a_k + b_k)}{\Gamma(a_k) \Gamma(b_k)} \int_0^1 u^a (1-u)^{b_k - 1} du$$

$$= \frac{a_k}{b_k} \quad (7)$$

$$\therefore H(x_f | t, data) = \frac{\beta G(x_f | t) + (k+1) F_k(x_f)}{\beta + K}$$

Equation (7) means that the predictive distribution is a weighted mean of the proposed distribution $G(x_f | t)$ and the empirical distribution $F_k(x_f)$.

2.2. Model Assessment

Let $p = \frac{i}{k+1}, i = 1, 2, \ldots, k$. Then by predicting k future values (quantiles), $Q(p)$, corresponding to probabilities p, we can compare the predicted values with the observed. The calculation of the predictive quantiles from $H(x|t, x_1, x_2, \ldots, x_k)$ is done numerically by letting $H = p, 0 < p < 1$ and solving for the p^{ith} quantile x_f numerically. There are different ways of solving for x_f. In this study, we solve for x_f by discretization of the sample space into small intervals and then choose a value for p according to these intervals.

To assess the fit of the model, we visually compare the predicted distribution with the observed values by a Q-Q plot of the predicted quantiles versus the observed values. If the model fits the data well, then its predictions should resemble the data, and we would expect the points to approximately follow a straight line; large discrepancies between the observed data and the predicted observations indicate poor fit.

To complement the Q-Q plot of the observed data and the predicted observations we use a numerical measure of goodness of fit, Nash-Sutcliffe model efficiency coefficient (due to [29]). The coefficient of efficiency measures the error in the predicted observations relative to the natural variation in the observed values, and is given by

$$E_{N-S} = 1 - \frac{\sum_{i=1}^{k} (Q_i - Q_i(p))^2}{\sum_{i=1}^{k} (Q_i - \overline{Q})^2} \quad (8)$$

where $Q_i(p)$ is the ith predicted quantile, Q_i is the ith observed value, for $i = 1, 2, \ldots, k$, and $\overline{Q}$ is the mean of the observed values. An $E_{N-S} = 1$ indicates perfect fit between the observed and predicted data (perfect model fit). An $E_{N-S} = 0$ indicates that the model is predicting no better than using the average of the observed data. Increasingly negative values of E_{N-S} indicate increasingly poorer predictions (poor model fit). See [30].

3. Application

In civil designs, extreme river discharges with very large return periods can be estimated by fitting various probability models to available observations. The need for a predictive hydrological model of the Okavango Delta was recognized a long time ago, and several modelling efforts have been carried out, [2, 18], presented a statistical model to predict the extent and geographical area of the annual maximum flooding of the Okavango Delta from observed inflow and local precipitation. In this paper, we model the annual

maximum water inflows into the Okavango River at Mohembo, Botswana, during the period 1975-2003, by using the Dirichlet process based generalized Pareto model. The data are given in Figure 1.
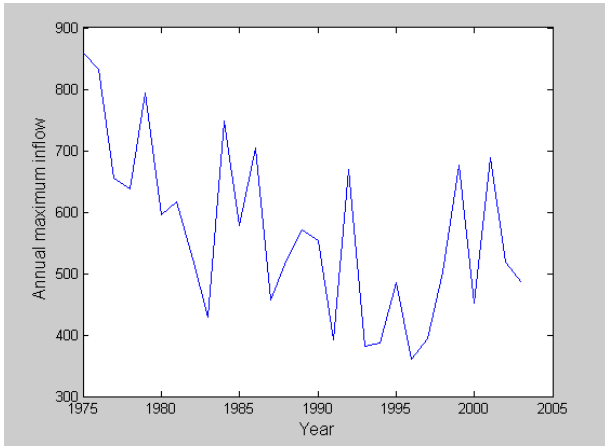

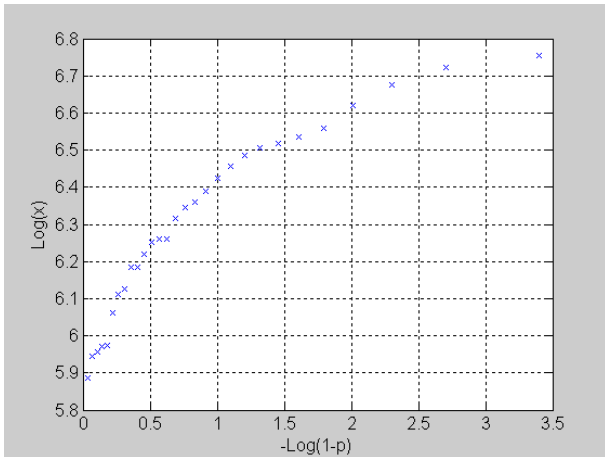**Figure 1.** *Annual maximum inflows (m³/sec) at Mohembo, Botswana, 1975-2003.*

Figure 2. *Pareto quantile plot of annual maximum inflow of at Mohembo.*

To check if the distribution that generated the inflow data is in the Frechet-Pareto domain of attraction, we graph a Pareto quantile plot of the annual maximum inflows, given in Figure 2. The Pareto quantile plot in Figure 2 is approximately linear, but bends down, at the very largest observations. This indicates a weaker behavior of the ultimate tail of the annual maximum inflows distribution. So we can reasonably conclude that that the underlying distribution of the annual maximum inflows of the Okavango river at Mohembo belongs in the Frechet-Pareto class, and, therefore, proceed to fit a generalized Pareto distribution $GPD(\gamma, \sigma_t)$ to the data above a certain threshold, t. The parameter γ denotes the extreme value index and determines the weight of the tail of G, and hence F, while σ_t is the scale parameter and is related to the variation in the data at the threshold t. These parameters are estimated here through Markov chain Monte Carlo simulation.

Recall the distribution function for $X|t \sim GPD(\gamma, \sigma_t)$ given by Equation (2). Assume the maximal data information

(MDI) prior [32] on the parameters (γ, σ_t), given by

$$\pi(\gamma, \sigma_t) \propto \frac{1}{\sigma_t} e^{-\gamma} \qquad (9)$$

From equations (2) and (9), we obtain the posterior density of the $GPD(\gamma, \sigma_t)$ as

$$\pi\left(\gamma, \sigma_t \middle| x_{(1)}, x_{(2)}, \ldots, x_{(k)}\right) \propto \frac{1}{\sigma_t^{k+1}} \prod_{i=1}^{k} \left(1 + \frac{\gamma\left(x_{(i)} - t\right)}{\sigma_t}\right)^{-\frac{1}{\gamma}-1} \qquad (10)$$

We use the Metropolis-Hastings algorithm with proposal density q taken as lognormal $\log(\gamma, \sigma_t) \sim N(\gamma, \sigma_t)$ to restrict γ to be positive. Initializing with $(\gamma, \sigma_t) = (0.213, 210)$, Figure 3 (a)-3 (b) shows the behavior of Bayes estimates of γ and σ_t for various values of k, the number of observations above threshold t, after 1000 iterations of the chain for each $k = 5, 6, \ldots, 29$.
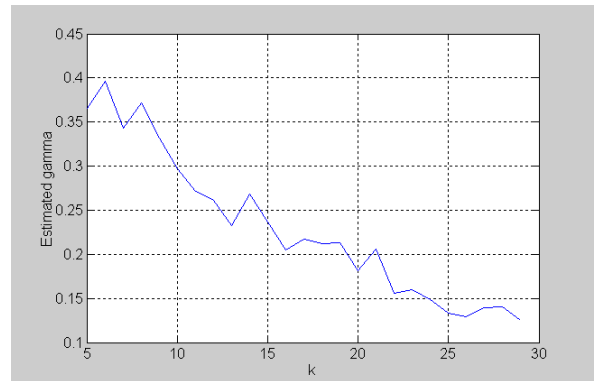

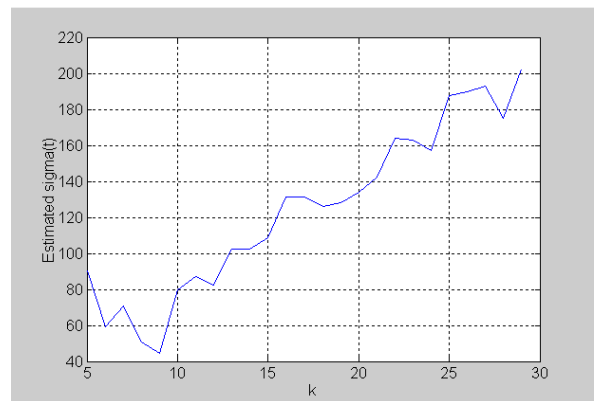**Figure 3.** *Estimates of gamma against k for annual maximum inflow data.*

Figure 4. *Estimated sigma (t) against k for annual maximum inflow data.*

Estimates of γ and σ_t are given by their respective posterior modes, producing Bayes estimates $\hat{\gamma} = 0.213$ and $\sigma_t = 130$, respectively. From Figure 3, we notice that estimates of γ stabilize around $\hat{\gamma} = 0.213$, while Figure 4 shows that estimates of σ_t for various k values are stable

within the range $16 \leq k \leq 20$, suggesting take $\sigma_t = 130$.

To select the threshold t through minimizing the negative differential entropy of the Dirichlet process, we need estimates of parameters of the Dirichlet process $v_i = \beta \left\{ G \left(x_{(i)} | t \right) - G \left(x_{(i-1)} | t \right) \right\}$, $i = 1, 2, \ldots k$. This implies that we also need an estimate of the concentration parameter, $\beta = \alpha(k+1)$, or equivalently, of α. The choice of α for a given k is obtained by minimizing the negative differential entropy of the Dirichlet distribution. This leads to various selected values of α corresponding to various minimal values of the negative differential entropy for different values of k. Thus, to find the optimal value of α, plot the selected α's against k as in Figure 4, and select as the optimal the alpha corresponding to a region within which the estimates are stable.

From Figure 4, we take $\hat{\alpha} = 0.26$, ignoring the first part of the graph below $k = 16$ where the estimates are quite unstable, and the last part of the graph after $k = 25$ where the threshold t would be too low. Too low a threshold is likely to violate the asymptotic basis of the model, leading to bias; too high a threshold will generate few excesses with which the model can be estimated, leading to high variance [8]. For this value of α, Figure 5 shows that the minimum value of the relative negative differential entropy is reached when $k = 16$ corresponding to a selected threshold of $t = 523.33$ m^3/sec.
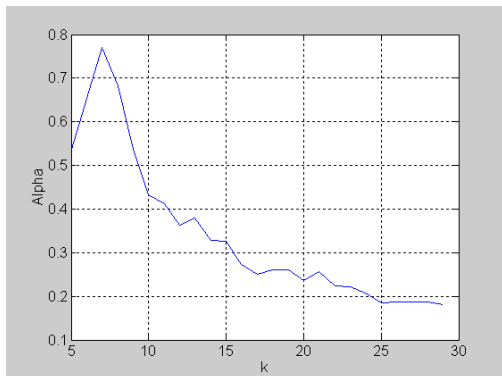

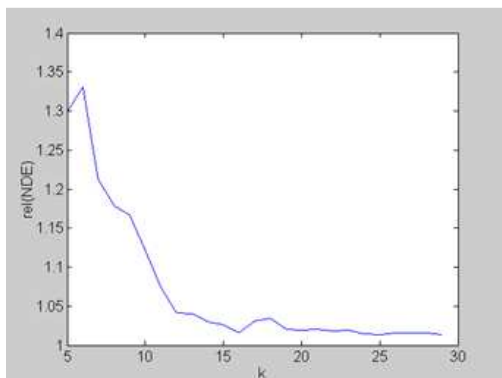**Figure 5.** *Selected values of alpha against k for annual maximum inflow data.*

Figure 6. *Relative negative differential entropy against k for annual maximum data.*

Using the results above, we fit a generalized Pareto distribution with $\hat{\gamma} = 0.213$ and $\sigma_t = 130$ to the $k = 16$ largest observations. The empirical distribution function, F_k, the proposed GPD model and the predictive (PD) distribution functions are given in Figure 7. Figure 7 shows that the GPD gives a good fit to the data in the very largest observations, but poor model fit in the lower largest observations. The posterior predictive distribution gives reasonably good overall fit to the extreme data above t. However, a Q-Q plot of the predictive quantiles versus the observed, hereafter referred to as a *posterior predictive quantile plot*, is more informative. See Figure 7. From Figure 8, both the predictive distribution (PD) and the generalized Pareto distribution quantiles (estimated) are consistent with the observed annual maximum inflow data for the Okavango River. Notice, however, that that the estimated GPD quantiles and the predicted quantiles differ at the very largest observations. This weaker behavior of the ultimate tail of the annual maximum inflows distribution was observed in the generalized Pareto quantile plot (Figure 2). Overall, the posterior predictive quantile plot shows a good agreement between the predicted values and the observed values, implying the generalized Pareto distribution fits the annual maximum inflow data reasonably well.
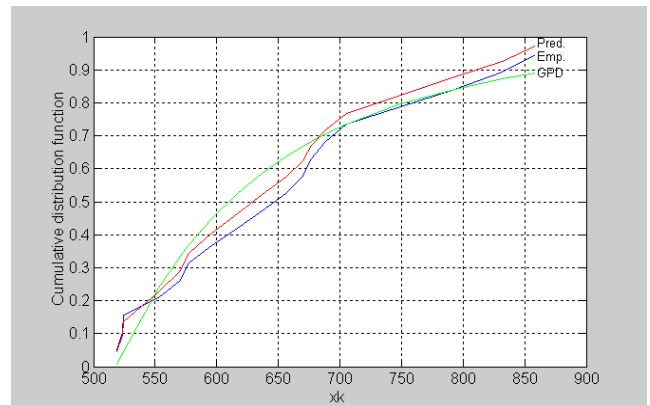

Figure 7. *Distribution functions of F, G, and H for annual maximum inflow data.*
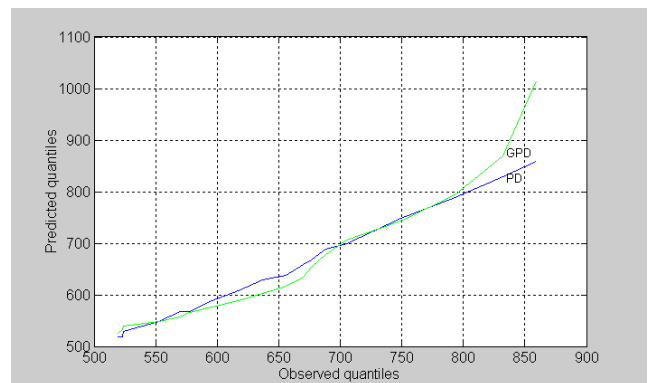

Figure 8. *Q-Q plots of predicted distribution (PD) and estimated GPD.*

The Nash-Sutcliffe coefficient of model efficiency for this dataset is found to be $E_{N-S} = 0.9952$ (very close to 1), indicating that the model produces predictions that closely

resemble the observed data. This value of a numerical measure of discrepancy between the predicted and the observed values further supports the evidence given graphically in terms of the posterior predictive quantile plot. Thus, we can conclude that the Dirichlet based generalized Pareto model fits the annual maximum inflow data very well.

4. Conclusions

The results from this investigation show that posterior predictive checks seem to be reasonable diagnostic tools for assessing the fit of the generalized Pareto. In particular, the posterior predictive quantile plot seems to be more informative in assessing the fit of an extreme value model than the probability plot. Selecting a threshold by minimizing the negative differential entropy of a Dirichlet process has the added advantage of allowing the analyst to estimate the concentration parameter from the model, rather than specifying it as a measure of his belief in the proposed model as a prior guess for the unknown distribution that generated the observations. Lastly, the results of the analysis of the annual maximum inflows data show that the distribution of the annual maximal inflows into the Okavango River at Mohembo, Botswana, can be adequately described by the generalized Pareto distribution.

References

[1] Ames, A. J. (2018). Prior Sensitivity of the Posterior Predictive Checks Method for Item Response Theory models. Vol. 16, NO. 4, 239-255.

[2] Bauer, P., Gumbricht, T. and Kinzelbach, W. (2006). A regional coupled surface water/groundwater model of the Okavango Delta, Botswana. Water resources Research, Vol. 42, W04403, doi: 10.1029/2005WR004234.

[3] Berger, J. O. and Guglielmi, A. (1999). Bayesian testing of a parametric model versus nonparametric alternatives. Worldwide website: http://www.isds.duke.edu/~berger/papers/99-04.html.

[4] Berger, J. O. and Guglielmi, A. (2001). Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. Journal of American Statistical Association 96, 174-184.

[5] Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004). Statistics of Extremes, Theory and Applications. Wiley.

[6] Box, G. E. P. and Tiao, G. C. (1973). Bayesian Inference in Statistical Analysis. New York. Wiley Classics.

[7] Coles, S. G. and Powel, E. A. (1996). Bayesian Methods in Extreme Value Modelling: A Review and New Developments. International Statistical Review, Vol. 64, No. 1, pp. 119-136.

[8] Coles, S. G. (2001). An Introduction to Statistical Modeling of Extreme Values. Springer, London.

[9] Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high threshold (with discussion). Journal of the Royal Statistical Society, B62, 393-442.

[10] De Waal, D, J. (2005). Predictive density derived from the Dirichlet process on goodness of fit of extreme modelling. Tech. Report Nr. 351. Department of Mathematical Statistics, University of the Free State.

[11] De Waal, D, J. and Beirlant, J. (2005). Choice of the threshold through the entropy of the Dirichlet Process when applying POT. Tech. Report nr. 351, Department of Mathematical Statistics, University of the Free State.

[12] De Waal, D, J., Beirlant, J. and Dierckx, G. (2008). Predicting high quantiles through the Dirichlet process on extreme modelling. South African Statistical Journal, Volume 42, Issue 2, Jan 2008, p. 101 – 124.

[13] Fawcett, L. and Walshaw, D. (2016). Sea-surge and wind speed extremes: optimal estimation strategies for planners and engineers. Stochastic Environmental Research and Risk Assessment 30: 463-480.

[14] Fawcett, L. and Walshaw, D. (2018). Bayesian posterior predictive return levels for environmental extremes. Stochastic Environmental Research and Risk Assessment 32: 2233-2252.

[15] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. Annals of Statistics Vol. 1, 209-230.

[16] Gelman, Carlin, Stern and Rubin, 2004. Bayesian Data Analysis. 2nd Ed. Chapman and Hall/CRC.

[17] Gelman, A., Meng, X. L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica 6, 733-807.

[18] Gumbricht, T., Wolski, P., Frost, P. and McCarthy, T. S. (2004). Forecasting the spatial extent of the annual flood of the Okavango Delta, Botswana. Journal of Hydrology 290, 178-191.

[19] Guttman, (1967). The use of the concept of a future observation in goodness-of-fit problems. Journal of the Royal Statistical Society B 29, 83-100.

[20] Honkela, A. (2001). Dirichlet distribution. Worldwide website: http://www.hut.fi/ahonkela/dippa/node95.html.

[21] Janssen, A (2003). Which power of goodness of fit tests can really be expected: intermediate versus contiguous alternatives. Statistics & Decisions 21, 301-325.

[22] Kapur, J. N. and Kesavan, H. K. (1992). Entropy Optimization Principles with Applications. Academic Press, Inc.

[23] Leadbetter, M. R. (1991). On a basis for 'Peaks over Threshold' modelling. Statistics and Probability Letters 12, 357-362.

[24] Ledwina, (1994). Data-driven version of Neyman's smooth test of fit. Journal of American Statistical Association 89, 1000-1005.

[25] Lynch, M. L. and Western, B. (2004). Bayesian Posterior Predictive Checks for complex models. Sociological methods & Research, Vol. 32, 301-335.

[26] Mazzuchi (2002). Bayes estimate and inferences for entropy and information index of fit. SANPAD Workshop Proceedings. TU Delft. Worldwide website: http://www.waterbouw.tudelft.nl/public/gelder/sanpadworksho p2002.pdf.

[27] Mimno, D., Blei, D. M. and Engelhardt, B. E. (2015). Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure. PNAS June 30, 2015 112 (26) E3441-E3450.

[28] Mothupi, T., Thupeng, W. M., Mashabe, B. and Mokoto, B. (2016). Estimating Extreme Quantiles of the Maximum Surface Air Temperatures for the Sir Seretse Khama International Airport Using the Generalized Extreme Value Distribution. American Journal of Theoretical and Applied Statistics 5 (6): 365-375.

[29] Nash, J. E. and Sutcliffe, J. V (1970). River floor forecasting through conceptual models Part 1 – A discussion of principles. Journal of Hydrology 10 (3), 282-290.

[30] Oztekin, T. (2005). Comparison of Parameter Estimation Methods for the Three-Parameter Generalized Pareto Distribution. Turk j of Agric For 29, 419-428.

[31] Pickands, J. (1975). Statistical inference using extreme order statistics. Annals of statistics 3, 119-131.

[32] Rubin, D. B. (1981). Estimation in parallel randomized experiments. Journal of Educational Statistics 6, 377-401.

[33] Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. Annals of Statistics 12, 1151-1172.

[34] Shannon, C. E., (1948). "A Mathematical Theory of Communication". Bell System Tech. J. 27: 379-423, 623-659.

[35] Smith, R. L. (1989). Extreme value analysis of environmental time series: an application to.

[36] Trend detection in grouped-level ozone, Statistical Science 4, 367-393.

[37] Verdinelli, I. and Wasserman, L. (1998). Bayesian goodness of fit testing using infinite dimensional exponential families. The Annals of Statistics 20, 1203-1221.

[38] Zellner, A. (1977). "Maximal Data Information Prior Distributions". In A. Aykac and C. Brumat, Eds. New Developments in the Applications of Bayesian Methods. Amsterdam, North Holland Publishing Co., 211-232.