Extremes manuscript No.
(will be inserted by the editor)

Improved threshold diagnostic plots for extreme value analyses

Paul J. Northrop · **Claire L. Coleman**

Received: date / Accepted: date

Abstract A crucial aspect of threshold-based extreme value analyses is the level at which the threshold is set. For a suitably high threshold asymptotic theory suggests that threshold excesses may be modelled by a generalized Pareto distribution. A common threshold diagnostic is a plot of estimates of the generalized Pareto shape parameter over a range of thresholds. The aim is to select the lowest threshold above which the estimates are judged to be approximately constant, taking into account sampling variability summarized by pointwise confidence intervals. This approach doesn't test directly the hypothesis that the underlying shape parameter is constant above a given threshold, but requires the user subjectively to combine information from many dependent estimates and confidence intervals. We develop tests of this hypothesis based on a multiple-threshold penultimate model that generalizes a two-threshold model proposed recently. One variant uses only the model fits from the traditional parameter stability plot. This is particularly beneficial when many datasets are analysed and enables assessment of the properties of the test on simulated data. We assess and illustrate these tests on river flow rate data and 72 series of significant wave heights.

Keywords Extreme value theory · Generalized Pareto distribution · Score test · Likelihood ratio test · Penultimate approximation · Threshold.

Mathematics Subject Classification (2000) 62G32 · 62F03 · 62F05 · 62P12 · 62P35

Paul J. Northrop
Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, UK
Tel.: +44-20-76791869
Fax: +44-20-31083105
E-mail: p.northrop@ucl.ac.uk

Claire L. Coleman
Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, UK
Present address: Pragmatic Clinical Trials Unit (PCTU), Centre for Primary Care and Public Health, Blizard Institute, Barts and The London School of Medicine & Dentistry, Yvonne Carter Building, 58 Turner Street, London E1 2AB, UK

1 Introduction

Extreme value theory provides asymptotic justification for particular families of models for extreme data. Let $X_1, X_2, \ldots X_n$ be a sequence of independent and identically distributed random variables and u_n a threshold, increasing with n. Pickands (1975) showed that if there is a non-degenerate limiting distribution for appropriately linearly rescaled excesses of u_n then this limit is a Generalized Pareto (GP) distribution. In practice, a suitably high threshold u is chosen empirically. Given that there is an exceedance of u, the excess $Y = X - u$ is modelled by a GP(σ_u, ξ) distribution, with positive threshold-dependent scale parameter σ_u, shape parameter ξ and distribution function

$$G(y) = \begin{cases} 1 - (1 + \xi y/\sigma_u)_+^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-y/\sigma_u), & \xi = 0, \end{cases} \tag{1}$$

where $y > 0$, $x_+ = \max(x, 0)$. The $\xi = 0$ case is defined in the limit as $\xi \to 0$. The number of exceedances is modelled by a Poisson distribution with threshold-dependent mean λ_u. An equivalent formulation (Pickands, 1971) is the non-homogeneous Poisson process (NHPP) representation, whose threshold-independent parameterization has advantages if, for example, covariates effects in the parameters are modelled.

We consider fixed threshold selection: choosing a single value of u to be treated as fixed and known when subsequent inferences are made. This involves a bias-variance trade-off: the lower the threshold the greater the estimation bias due to model misspecification; the higher the threshold the greater the estimation uncertainty. Scarrott and MacDonald (2012) provides a comprehensive review of threshold selection approaches.

We consider a widely-used informal graphical approach (described in section 1.1) based on the property that ξ should be constant over thresholds for which the GP (or NHPP) model is appropriate. Wadsworth and Tawn (2012) formalize this approach using a NHPP likelihood-based test of whether sufficient convergence has been achieved. However, the test uses a very computationally-intensive simulation scheme, which limits its practical usefulness. Working within the Wadsworth–Tawn paradigm we develop a substantially faster test by avoiding the need for simulation.

1.1 Traditional parameter stability plots

These plots examine the sensitivity of GP parameter estimates to the threshold and are described in Coles (2001, chapter 4) and Scarrott and MacDonald (2012). Over a range of values of u, estimates $\widehat{\xi}$ of the shape and $\widehat{\sigma}_u - \widehat{\xi} u$ of the modified scale are plotted against u with pointwise 95% symmetric confidence intervals. The lowest threshold above which these quantities are judged to be approximately constant in u is selected, taking into account sampling variability summarized by the confidence intervals.

For comparability with Wadsworth and Tawn (2012) we take as an example the set of 154 flow rates exceeding 65 $\mathrm{m}^3\mathrm{s}^{-1}$, over the period 1934–1969, from the River Nidd in Yorkshire, UK. These data have been processed so that the data can be treated

as independent (Davison and Smith, 1990). Figure 1 shows parameter stability plots based on these data. As the threshold varies the estimates of the shape and modified

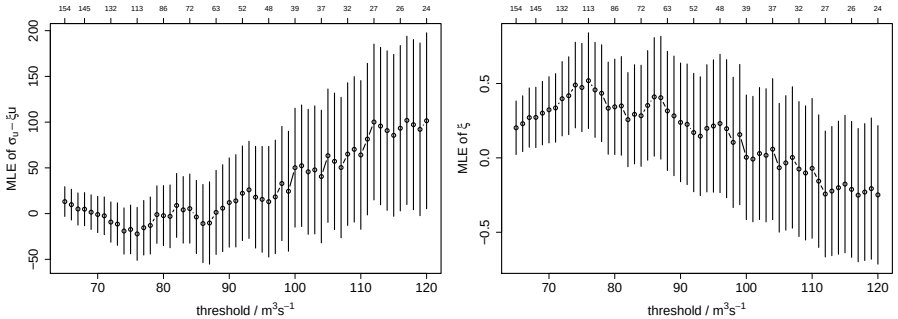

Fig. 1 Nidd flow rate example. Parameter stability plots over a range of thresholds, with 95% pointwise symmetric confidence intervals. Left: modified scale parameter. Right: shape parameter. The upper axis scale gives the number of threshold exceedances.

scale parameters move almost in direct opposition to each other, as is typical in these plots (Scarrott and MacDonald, 2012). Thus, the modified scale plot, for example, is somewhat unnecessary. It is expected that the parameter estimates become unstable at very high thresholds: for small sizes there is large sampling variability and the movement of the threshold through the data points can result in a lack of smoothness.

Selecting a threshold based on the right hand plot in figure 1 is problematic: the estimates at two different thresholds are strongly dependent; the viewer will compare many pairs of thresholds, so there is a multiple testing issue; and looking for whether or not confidence intervals overlap does not make the desired comparison appropriately. One really wants to test the null hypothesis $H_0 : \xi(u) = \xi(u_0)$, for all $u \geqslant u_0$, for some u_0, where $\xi(u)$ is the underlying value of the shape parameter at threshold u. In the next section we consider an approximate model for the subasymptotic behaviour of the shape parameter under which a discretized version of this hypothesis can be tested more objectively.

1.2 A piecewise constant shape parameter approximation

Wadsworth and Tawn (2012) propose a penultimate NHPP model in which the shape parameter is modelled as a piecewise constant function of threshold. This is motivated by theory suggesting that the structural form of the NHPP model (and the GP model for threshold excesses) holds well for relatively low thresholds with a shape parameter that is a slowly changing function of threshold. For a pair of thresholds (u,v), where $v < u$, the shape parameter has a change-point at u:

$$\xi(x) = \begin{cases} \xi_{vu}, & v < x < u, \\ \xi_u, & x > u. \end{cases} \tag{2}$$

This representation is used for two purposes: fixed threshold selection using frequentist tests of $\xi_{vu} = \xi_u$ for all pairs of values for u and v from a set of thresholds and a Bayesian assessment of threshold uncertainty, in which v is fixed and u is treated as a parameter. For the former the multiple-testing issue is solved using simulation to test the null hypothesis of equality of shape parameter at and above some lowest threshold $v_{\min}$. The computational intensity of the simulation scheme is an obstacle to the use of this model for fixed threshold selection. To tackle the multiple-testing issue directly, and (as suggested on page 552 of Wadsworth and Tawn (2012)) to attain a better approximation to $\xi(x)$, we extend the piecewise constant representation to an arbitrary number m of thresholds $(u_1, \ldots, u_m)$:

$$\xi(x) = \begin{cases} \xi_i, & u_i < x < u_{i+1}, \quad \text{for } i = 1, \ldots, m-1, \\ \xi_m, & x > u_m. \end{cases} \tag{3}$$

For a given set of thresholds a single test based on (3) is no more demanding of the data than the combination of all pairwise tests based on (2): the latter involves making inferences on each interval $(u_i, u_{i+1}), i = 1, \ldots, m$. Also, (2) assumes constancy of the shape parameter over wider intervals than (3).

In section 2 we define a multiple-threshold Generalized Pareto model based on (3) and two tests, a likelihood ratio test and a score test, for stability in the shape parameter of the model with threshold. In section 3 we check that the most computationally efficient of these tests performs as expected on simulated data. In section 4 we present analyses of the River Nidd dataset and, to illustrate the potential of our approach for use on multiple datasets, of a set of 72 series of significant wave heights from the Gulf of Mexico. In the appendix we derive a Fisher expected information matrix for use in the score test. Computer code to implement this methodology is available at `www.homepages.ucl.ac.uk/~ucakpjn/`.

2 A multiple-threshold GP model

Consider m thresholds $u_1 < u_2 < \cdots < u_m$. Let $v_j = u_j - u_1$, for $j = 1, \ldots, m$ and $w_j = u_{j+1} - u_j = v_{j+1} - v_j$, for $j = 1, \ldots, m-1$. Let Y denote an excess of u_1. For $j = 1, \ldots, m$ we assume that $(Y - v_j) \mid v_j < Y < v_{j+1}$ has a (truncated) GP(σ_j, ξ_j) distribution, where $v_{m+1} = v_m - \sigma_m/\xi_m$ if $\xi_m < 0$ and otherwise v_{m+1} is infinite. The conditional density of $(Y - v_j) \mid v_j < Y < v_{j+1}$ is given by

$$f_{(Y-v_j) \mid v_j<Y<v_{j+1}}(y - v_j) = \frac{f_j(y - v_j)}{F_j(w_j)}, \quad 0 < y - v_j < w_j,$$

where $f_j(y - v_j) = \sigma_j^{-1}[1 + \xi_j(y - v_j)/\sigma_j]_+^{-(1+1/\xi_j)}$ is a GP(σ_j, ξ_j) density function for $Y - v_j$ and $F_j(w_j) = 1 - [1 + \xi_j w_j/\sigma_j]_+^{-1/\xi_j}$ normalises the conditional density. Let $p_j = P(Y > v_j)$. Then $p_1 = 1$ and

$$p_j = \prod_{i=1}^{j-1} [1 - F_i(w_i)] = \prod_{i=1}^{j-1} \left[1 + \frac{\xi_i w_i}{\sigma_i}\right]_+^{-1/\xi_i}, \quad j = 2, \ldots, m.$$

Therefore, $P(v_j < Y < v_{j+1}) = p_j - p_{j+1} = p_j F_j(w_j)$ and an equivalent formulation is

$$f(y) = \prod_{j=1}^{m} \{p_j f_j(y)\}^{I_j}, \tag{4}$$

where $I_j = I(v_j < y < v_{j+1})$ and $I(A)$ is the indicator function of the set A. Thus, the shape parameter is modelled as a piecewise constant function $\xi(y)$ with change-points at $v_j, j = 2,\ldots,m$. In order that there is no discontinuity in $f(y)$ we set $\sigma_{j+1} = \sigma_j + \xi_j w_j$, for $j = 1,\ldots,m-1$, so that $\sigma_j = \sigma_1 + \sum_{i=1}^{j-1} \xi_i w_i$. The parameters of the model are $\psi = (\sigma_1, \xi_1, \ldots, \xi_m)$, where $\sigma_1 > 0$.

For a random sample $y = (y_1,\ldots,y_n)$ of excesses of u_1 from density (4) the log-likelihood is

$$l(\sigma_1, \xi_1, \ldots, \xi_m) = \sum_{i=1}^{n} \sum_{j=1}^{m} I_{ij} \left\{ \log p_j - \log \sigma_j - \left(1 + \frac{1}{\xi_j}\right) \log \left[1 + \frac{\xi_j (y_i - v_j)}{\sigma_j}\right] \right\}, \tag{5}$$

where $I_{ij} = I(v_j < y_i < v_{j+1})$.

2.1 Likelihood ratio and score tests

Consider threshold u_1. We wish to assess whether a common GP model applies on all intervals $(v_k, v_{k+1}), k = 1,\ldots,m$. That is, we wish to test $H_0 : \xi_1 = \cdots = \xi_m$. Rejection of H_0 suggests that a threshold higher than u_1 is required.

Let $\tilde{\sigma}_1$ and $\tilde{\xi}_1$ denote the (restricted) MLEs of σ_1 and ξ_1 under the null hypothesis, that is, from a GP fit to excesses of u_1. Let $\widehat{\sigma}_1, \widehat{\xi}_i, i = 1,\ldots,m$ denote the (unrestricted) MLEs. In the appendix expressions for the score function and Fisher expected information are derived, under the parameterization $\theta = (\theta_1,\ldots,\theta_{m+1}) = (\sigma_1, \phi_1, \ldots, \phi_m)$, where $\phi_j = \xi_j/\sigma_j$. Let $\widehat{\theta}_0$ and $\widehat{\theta}$ denote the restricted and unrestricted MLEs of θ.

We consider the likelihood ratio (LR) and score test statistics

$$W = 2\left\{ l(\widehat{\theta}) - l(\widehat{\theta}_0) \right\}, \tag{6}$$

$$S = U(\widehat{\theta}_0)^T \, i^{-1}(\widehat{\theta}_0) \, U(\widehat{\theta}_0), \tag{7}$$

where $U(\theta)$ is the score function and $i(\theta)$ is the expected information matrix. Provided that $\xi_m > -1/2$ (Smith, 1985) in each case the asymptotic null distribution of the statistic is χ^2_{m-1}. As the motivation for considering the score test is to avoid fitting the full model we use the expected information evaluated at the restricted MLE, i.e. under the null hypothesis. The expected information is used because positivity of the test statistic and consistency of the test are ensured, properties not achieved by the observed information (Freedman, 2007; Morgan et al, 2007). Moreover, in the current context, the observed information is known to have poor finite-sample properties (Süveges and Davison, 2010). The statistic S has the advantage over W that it requires only a fit of the null model at the threshold of interest. In contrast calculation of W is more difficult and time-consuming as it requires m shape parameters to be estimated. When m is large convergence problems can occur in the search for $\widehat{\theta}$, for

which we use $\widehat{\theta}_0$ as an initial estimate. We have not experienced this problem for any of the real datasets we have studied but it has occurred for some simulated datasets. It is possible that convergence could be achieved with better initial estimates but we expect that finding a general strategy for choosing such estimates is difficult.

Suppose now that the lowest threshold considered is u_i, so that the set of thresholds is $(u_i, \ldots, u_m)$. The null hypothesis is $H_0 : \xi_i = \cdots = \xi_m$ and the asymptotic null distribution of each test statistic is χ^2_{m-i}, providing, for $i = 1, \ldots, m-1$, p-values associated with the test of whether a threshold higher than u_i is required. Provided that the chi-squared distributions provide reasonable approximation to the null distributions no simulation is required. The score test is particularly efficient computationally as it uses only the (first $m-1$) model fits from the traditional parameter stability plot.

3 Simulation studies

We check the basic properties of the score test on random samples simulated from the unit exponential distribution and the standard normal distribution. For the former, excesses of any non-negative threshold are unit exponential (GP with shape 0), so H_0 is true for any such threshold. For the latter, H_0 is not true at any threshold but could be considered as becoming closer to the truth as the threshold increases. We also consider a distribution defined so that H_0 is false below a certain threshold and true otherwise. We use a hybrid distribution on the positive real line, with a constant density up to its 75% quantile and a GP density with shape parameter 0.1 for excesses of the 75% quantile. Thus, H_0 is true from the 75% quantile upwards.

We use a set of $m = 10$ thresholds, from the respective median of the underlying distributions to the 95% quantiles, in steps of 5%. For a threshold set at the q% quantile the sample size is $N_q = N(1 - q/100)$. We use $N = 2000$ and $N = 500$, so that there are respectively 100 and 25 excesses of the highest threshold. We run 1000 simulations. For each threshold (excluding the 95% quantile) we calculate the proportion of simulations for which the test (of size 0.05) of the null hypothesis of stability of shape parameter above the threshold is rejected. In the plot on the left of figure 2 this proportion is plotted against q. If the test is well-calibrated the proportion should be close to 0.05 whenever H_0 is true (all thresholds for the exponential example and all threshold from the 75% quantile for the hybrid example) and this is indeed the case. Otherwise, the proportion of null rejections decreases with N_q and therefore decreases as q increases.

The plot on the right of figure 2 shows, for $N = 2000$, how the proportion of null rejections varies with m. For $m = 6$ the thresholds are at the $(50, 60, 70, 80, 90, 95)$% quantiles; for $m = 19$ they range from the median to the 95% quantile in steps of 2.5%. When H_0 is not true this proportion is larger for smaller values of m. This is because the test seeks to detect differences in shape parameter at different thresholds. With a small number of widely-separated thresholds these differences are larger and are detected with greater probability. The same findings apply for $N = 500$. We would not choose a small value of m in practice because our aim is to use a set of thresholds that approximates the range of possible thresholds considered. We have only been able to examine the properties of the LR test in a limited way, for the reasons given in

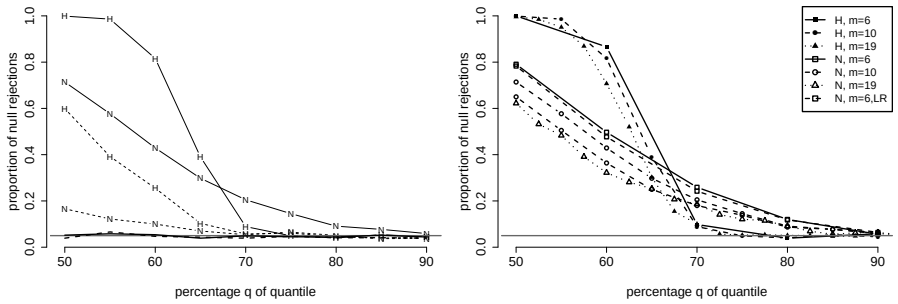

Fig. 2 Left: proportion of simulations for which the null hypothesis is rejected, based on a score test of size 0.05, against threshold for $N = 2000$ (solid lines) and $N = 500$ (dashed lines). The unlabelled lines relate to the exponential distribution. Lines for the normal and uniform-GP hybrid distributions are labelled with N and H respectively. A set of $m = 10$ thresholds is used. Right: proportion of null rejections for different values of m and $N = 2000$ for the normal and uniform-GP hybrid distributions. One line relates to the LR test. The horizontal grey lines are at 0.05.

section 2.1. The plot on the right in figure 2 includes a curve for the LR test for $m = 6$ based on simulations from the normal distribution. The proportion of null rejections is marginally smaller for the LR test than for the score test. For $m = 10$ the curve for LR test (not shown in the plot) lies further below that of the score test: it is only slightly above the score test curve for $m = 19$. A possible explanation is that sometimes local optima are found in the unrestricted optimisation of the log-likelihood and that this occurs more often for large m than for small m.

We extend the simulations from the exponential distribution to study the approximation of the null distribution of the score statistic S by the asymptotic chi-squared result. We replace the $m = 6$ case with $m = 4$ (thresholds at the $(50, 70, 90, 95)\%$ quantiles) in order to see more clearly the effect of m and run $n_{\text{sim}} = 10,000$ simulations in order to estimate with precision high quantiles of the null distribution of S. Otherwise, the setup is the same as above. We produce probability plots, that is, we plot $\{\left(j/(n_{\text{sim}}+1), H(s_{(j)})\right), j = 1, \ldots, n_{\text{sim}}\}$, where H is a chi-squared distribution function with the appropriate degrees of freedom. Figure 3 shows probability plots for $N \in \{300, 1000\}$ and $q \in \{50, 90\}$ for different values of m, truncated to a range $(0.9, 1)$ of practical interest. In the plots for $q = 90$ the lines for $m = 4$ and $m = 10$ are coincident because the same (two) thresholds are involved. Points below the line of equality indicate that the score test is conservative. As one would expect the conservatism increases with q and decreases with N. The empirical null distribution agrees better with the asymptotic result for large m than for small m, that is, the test is less conservative for large m. Under the null there are no differences in shape parameter across thresholds, but with a small number of thresholds one has fewer opportunities to achieve the nominal false positive rate than with a larger number of thresholds.

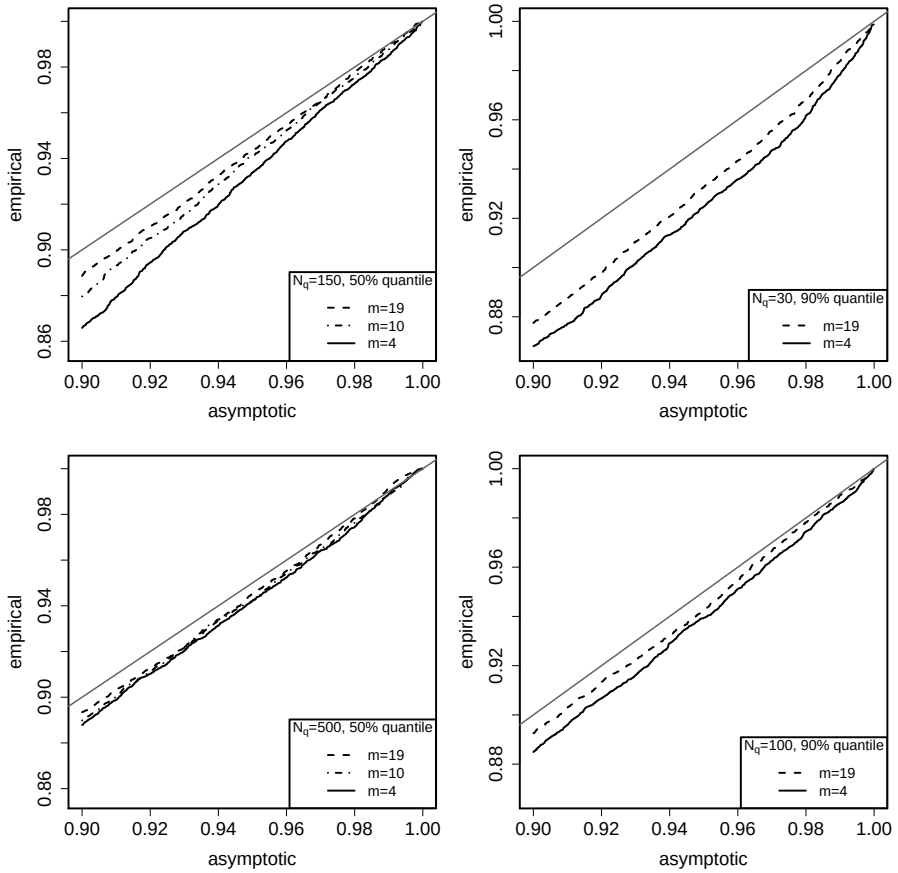

Fig. 3 Probability plots comparing the empirical null distribution of the score statistic to the asymptotic chi-squared result, where the null hypothesis is commonality of shape parameter above the $q\%$ quantile. Top row: sample size $N = 300$; $q = 50$ (left), $q = 90$ (right). Bottom row: $N = 1000$; $q = 50$ (left), $q = 90$ (right). N_q is the number of exceedances of the $q\%$ threshold.

4 Examples

We illustrate the multiple threshold diagnostic using two case studies. We return to the River Nidd data introduced in section 1.1. We also consider 72 time series of storm peak significant wave height on a spatial grid (Northrop and Jonathan, 2011), seeking to select a threshold for each of the 72 datasets. It is in such a situation, threshold selection for multiple datasets, that speed of computation of the multiple-threshold diagnostic is especially useful, particularly if the score test is used.

4.1 River Nidd flow rates

We use $m = 10$ thresholds from 65 to 120 in steps of 5. The plot on the left in figure 4 shows, for $i = 1,\ldots,m-1$, how the p-value associated with testing the null hypothesis $\xi_i = \cdots = \xi_m$ varies with the lowest threshold u_i. We calculate p-values using the score test and the likelihood ratio test based on the χ^2_{m-i} null distributions. We check the p-values using simulation. Consider threshold u_1 as an example. We simulate 1000 datasets from the GP$(\tilde{\sigma}_1, \tilde{\xi})$ model fitted under the null. For each simulated dataset we calculate the LR and score test statistics. For each test the estimated p-value is the proportion of the simulated test statistics that are greater than the corresponding observed test statistic. There is close agreement between the p-values based

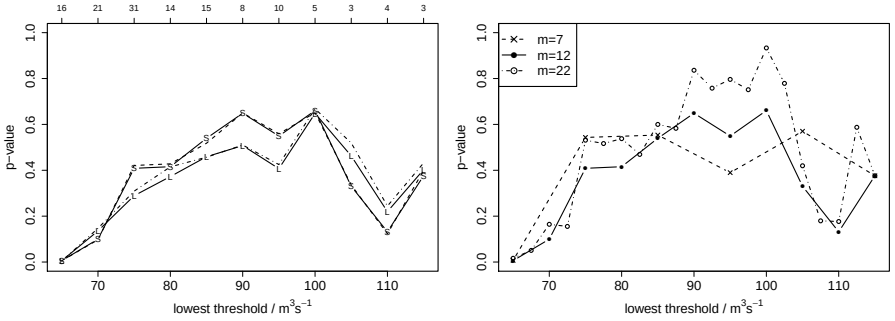

Fig. 4 Nidd flow rate example. Left: multiple threshold diagnostic plot for the threshold set $(u_1,\ldots,u_m) = \{65,70,\ldots,120\}$. p-values testing $\xi_i = \cdots = \xi_m$ plotted against u_i, for $i = 1,\ldots,m-1$: score test, using χ^2_{m-i} distribution (——S——); score test, using simulation $(- - -)$; LR test, using χ^2_{m-i} distribution (——L——); LR test, using simulation $(- \cdot - \cdot -)$. The upper axis scale gives the number of threshold exceedances between the thresholds for $m = 22$. There are 24 exceedances of 120 $\mathrm{m^3 s^{-1}}$. Right: p-values against u_i for different values of m.

on the χ^2_{m-i} distributions and on the simulations. The plot on the right in figure 4 examines how the choice of m affects the p-values. The general picture is similar over the different values of m but there are differences: the larger values of m pick up more of the location variation in parameter estimates evident in figure 1.

The threshold selected depends on how one wishes to make use of the p-values, and is subject to checking the fit of the GP model at this threshold. One could carry out a test of a given size (here using 5% suggests a threshold of 70 $\mathrm{m^3 s^{-1}}$) or view the p-values as a measure of the disagreement between the data and the null hypothesis. We discuss possibilities for automatic implementation of the former in section 4.2. For the latter, an idealised scenario is that the p-value increases with threshold until approximate stabilisation at a point where one could choose to set the threshold. Based on figure 4 one might choose a threshold of 75 $\mathrm{m^3 s^{-1}}$.

The results in Wadsworth and Tawn (2012), which includes tests based on the goodness-of-fit of the GP distribution across the thresholds, suggest that a threshold of 70 $\mathrm{m^3 s^{-1}}$ is too low. Figure 5 compares the GP fits at thresholds 70 $\mathrm{m^3 s^{-1}}$ and 75 $\mathrm{m^3 s^{-1}}$. For a threshold of 70 $\mathrm{m^3 s^{-1}}$, $\widehat{\xi} = 0.32$, with a 95% confidence interval

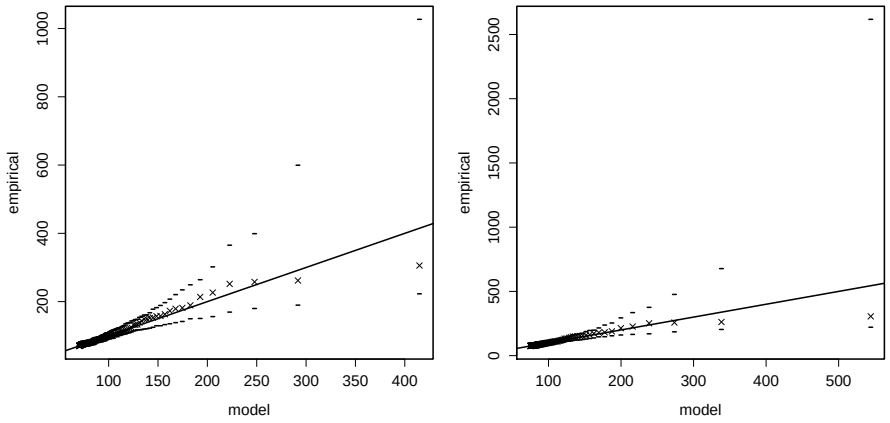

Fig. 5 Nidd flow rate example. Generalized Pareto Q-Q plot for the model fitted using thresholds of 70 $\mathrm{m^3 s^{-1}}$ (left) and 75 $\mathrm{m^3 s^{-1}}$ (right), with 95% simulation envelopes. A line of equality is superimposed.

of $(0.13, 0.58)$ and for a threshold of 75 $\mathrm{m^3 s^{-1}}$, $\widehat{\xi} = 0.47$, with a 95% confidence interval of $(0.22, 0.82)$.

Finally, we note that figure 4 shows similar features to the posterior distributions of threshold displayed in figure 8 of Wadsworth and Tawn (2012). If one allows a threshold as low as 65 $\mathrm{m^3 s^{-1}}$ then a threshold of approximately 75 $\mathrm{m^3 s^{-1}}$ is indicated, but if such a low threshold is ruled out then the situation is far less clear. For example, if the lowest threshold considered is 100 $\mathrm{m^3 s^{-1}}$ then a threshold close to this lower bound has greatest support, but not substantially greater support than higher thresholds.

4.2 Gulf of Mexico significant wave heights

We consider 72 time series of storm peak significant wave height (H_s^{sp}) from the Gulf of Mexico, analysed by Northrop and Jonathan (2011). The series contain a peak hindcast value of significant wave height from each of the 315 hurricane-induced storms that hit the area of interest over the period September 1900 to September 2005 (Oceanweather Inc., 2005). It is reasonable to treat the data from separate storms as independent, although the series are strongly spatially dependent over the 6 by 12 spatial grid of sites. To account for spatial non-stationarity we use sets of thresholds equal to the at-site empirical deciles $(0, 10, \ldots, 90)\%$.

Figure 6 shows how the p-values vary with the empirical decile at which the threshold is set. There is large variability in behaviour between the sites. For some sites the p-value rises quickly and remains high. For others (e.g. site 32) the p-value remains low until the sample median and then rises quickly. The p-value at site 17 rises above 0.05 but drops briefly back below 0.05 when the sample median is used as a threshold. Site 64 is an example where the p-value fluctuates over the higher thresholds.

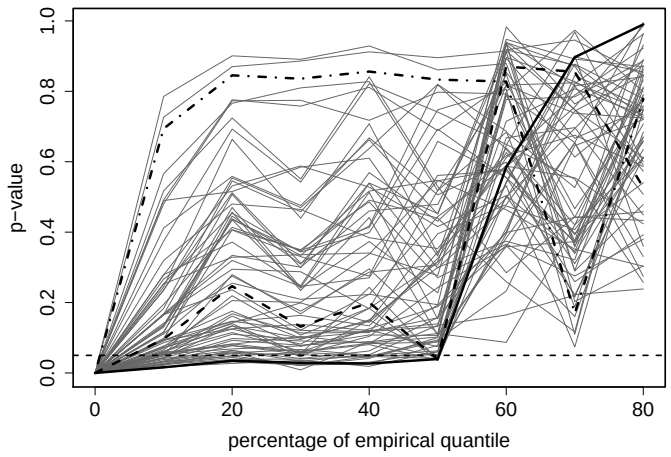

Fig. 6 Gulf of Mexico significant wave height example. At-site multiple threshold diagnostic plot for the threshold set $(u_1,\ldots,u_m) = (0,10,\ldots,90)\%$ empirical quantiles, using the score test. p-values testing $\xi_i = \cdots = \xi_m$ plotted against proportion of non-exceedances, for $i = 1,\ldots,m-1$: site 32 (——); site 17 $(- - -)$; site 64 $(- \cdot - \cdot -)$. The horizontal line is at 0.05.

Suppose that we wish to automate the selection of a threshold for each of the datasets, based on tests of size 5%, say. One possibility (Wadsworth and Tawn, 2012) is to select the lowest threshold for which the null hypothesis is not rejected. However, the p-values are not constrained to be non-decreasing in the lowest threshold. Therefore we might select the lowest threshold with the property that the null hypothesis is not rejected at it and at all the higher thresholds considered, bearing in mind that large variability is expected at the highest thresholds. Figure 7 shows the results of applying these two rules at each site in the significant wave heights dataset. In both cases thresholds at higher quantiles are required near the centre of the region than near the edges, with the highest threshold indicated being the 60% quantile. The latter is in agreement with Northrop and Jonathan (2011) who use quantile regression on latitude and longitude to set thresholds at estimated conditional quantiles of H_s^{sp}, judging that estimates of extreme value model parameters stabilize approximately around the 60% conditional quantile.

5 Discussion

We have developed a new approach to fixed threshold selection for extreme value models. Our approach extends the two-threshold penultimate extreme value model of Wadsworth and Tawn (2012) to an arbitrary number of thresholds, providing a better approximation to the subasymptotic behaviour of the generalized Pareto shape parameter. We implement tests of stability in shape parameter using likelihood ratio tests and score tests. For the latter we derive the Fisher information for the multiple-threshold model.

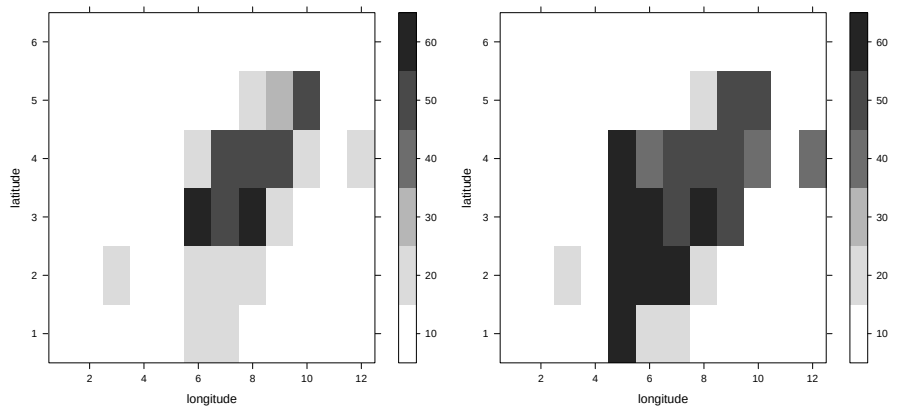

Fig. 7 Gulf of Mexico significant wave height example. Left: quantile after which p-value first rises above 0.05. Right: quantile after which p-value remains above 0.05.

Tests based on the multiple-threshold model are far more efficient computationally than repeated use of the two-threshold model over a grid of thresholds (for which a simulation scheme is required to adjust for multiple testing) and makes no greater demands on the data than the union of all these two-threshold fits. The speed of the score test is particularly useful for application to multiple datasets as it uses only the model fits from the traditional parameter stability plot. For example, producing the solid lines in figure 4 took 0.5 CPU seconds for the score test and 23.72 CPU seconds for the likelihood ratio test. Producing all 72 lines in figure 6 (using the score test) took 34.29 CPU seconds. A drawback of the fixed threshold approach is that it ignores the, perhaps considerable, uncertainty associated with the choice of threshold. Wadsworth and Tawn (2012) use tests of parameter stability to inform a Bayesian assessment of threshold uncertainty: the tests developed in this paper can speed up this process considerably.

We have assumed that threshold exceedances are independent. For the data we have considered this is reasonable because the raw data have been 'declustered': clusters of exceedances have been identified and only cluster maxima have been retained. More generally the GP model is an appropriate marginal model for threshold excesses (Anderson, 1990) and it is common for consideration of the effects of serial dependence to be ignored for threshold selection, entering only after the threshold is set (e.g. Fawcett and Walshaw (2012)). We expect that ignoring serial dependence results in tests that are liberal (i.e. adjustment for positive serial dependence would inflate the p-value) and ultimately selection of a higher threshold than is necessary. Further research will investigate the extent to which this is the case and how best to adjust p-values.

Acknowledgements We thank Yvo Pokern for comments on the traditional parameter stability plots that prompted this work. We are very grateful to two anonymous referees: their comments and suggestions have improved the original manuscript.

Appendix

We derive the score function u_1 and the Fisher expected information matrix i_1 for a single observation Y from density (4). We assume that $\xi_m > -1/2$ so that the likelihood satisfies the regularity conditions of Smith (1985). For convenience we work with the parameters $(\theta_1,\ldots,\theta_{m+1}) = (\sigma_1,\phi_1,\ldots,\phi_m)$, where $\phi_j = \xi_j/\sigma_j$. The log-likelihood is given by

$$l = \sum_{j=1}^{m} I_j \left\{ \log p_j - \log \sigma_j - \left(1 + \frac{1}{\sigma_j \phi_j}\right) \log\left[1 + \phi_j (Y - v_j)\right] \right\},$$

where $p_1 = 1$, $\log p_j = -\sum_{i=1}^{j-1} \log\left(1 + \phi_i w_i\right)/\sigma_i\phi_i$ and $\log \sigma_j = \log \sigma_1 + \sum_{i=1}^{j-1} \log(1 + \phi_i w_i)$, for $j = 2,\ldots,m$. In the following all empty products are 1 and all empty sums are 0.

Let $S_j = Y - v_j$, $T_j = 1 + \phi_j S_j$, $\gamma_j = 1 + \phi_j w_j$, $h_j = \phi_j \prod_{i=1}^{j-1} \gamma_i$ and $R_j = \sigma_j^{-1}\phi_j^{-1}\log T_j = \sigma_1^{-1} h_j^{-1} \log T_j$. The score vector is given by

$$\begin{aligned}
\frac{\partial l}{\partial \sigma_1} &= -\sigma_1^{-1} \sum_{j=1}^{m} I_j \left\{ \log p_j + 1 - R_j \right\}, \\
\frac{\partial l}{\partial \phi_k} &= \sum_{j=1}^{m} I_j \left\{ \frac{\partial \log p_j}{\partial \phi_k} - \frac{\partial \log \sigma_j}{\partial \phi_k} - \frac{\partial \log T_j}{\partial \phi_k} - \frac{\partial R_j}{\partial \phi_k} \right\},
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial \log \sigma_j}{\partial \phi_k} &= w_k \gamma_k^{-1} I(k \leqslant j-1), \\
\frac{\partial \log T_j}{\partial \phi_k} &= S_j T_j^{-1} I(k = j), \\
\frac{\partial R_j}{\partial \phi_k} &= \sigma_1^{-1} \left\{ h_j^{-1} \left[S_j T_j^{-1} - \phi_j^{-1} \log T_j \right] I(k = j) - w_k \gamma_k^{-1} I(k \leqslant j-1) h_j^{-1} \log T_j \right\}, \\
\frac{\partial \log p_j}{\partial \phi_k} &= -\sigma_1^{-1} \left\{ h_k^{-1} \left[w_k \gamma_k^{-1} - \phi_k^{-1} \log \gamma_k \right] I(k \leqslant j-1) - w_k \gamma_k^{-1} \sum_{i=1}^{j-1} I(k \leqslant i-1) h_i^{-1} \log \gamma_i \right\}.
\end{aligned}$$

The elements of the observed information matrix are given by

$$\begin{aligned}
-\frac{\partial^2 l}{\partial \sigma_1^2} &= -\sigma_1^{-2} \sum_{j=1}^{m} I_j \left\{ 2 \log p_j + 1 - 2\sigma_1^{-1} h_j^{-1} \log T_j \right\}, \\
-\frac{\partial^2 l}{\partial \sigma_1 \partial \phi_k} &= \sigma_1^{-1} \sum_{j=1}^{m} I_j \left\{ \frac{\partial \log p_j}{\partial \phi_k} - \frac{\partial R_j}{\partial \phi_k} \right\}, \\
-\frac{\partial^2 l}{\partial \phi_k^2} &= -\sum_{j=1}^{m} I_j \left\{ \frac{\partial^2 \log p_j}{\partial \phi_k^2} - \frac{\partial^2 \log \sigma_j}{\partial \phi_k^2} - \frac{\partial^2 \log T_j}{\partial \phi_k^2} - \frac{\partial^2 R_j}{\partial \phi_k^2} \right\},
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial^2 \log \sigma_j}{\partial \phi_k^2} &= -w_k^2 \gamma_k^{-2} I(k \leqslant j-1), \\
\frac{\partial^2 \log T_j}{\partial \phi_k^2} &= -S_j^2 T_j^{-2} I(k=j), \\
\frac{\partial^2 R_j}{\partial \phi_k^2} &= \sigma_1^{-1} \Bigg\{ h_j^{-1} \left[2\phi_j^{-2} \log T_j - 2\phi_j^{-1} S_j T_j^{-1} - S_j^2 T_j^{-2} \right] I(k=j) \\
&\qquad + 2w_k^2 \gamma_k^{-2} I(k \leqslant j-1) h_j^{-1} \log T_j \Bigg\}, \\
\frac{\partial^2 \log p_j}{\partial \phi_k^2} &= -\sigma_1^{-1} \Bigg\{ h_k^{-1} \left[2\phi_k^{-2} \log \gamma_k - 2\phi_k^{-1} w_k \gamma_k^{-1} - w_k^2 \gamma_k^{-2} \right] I(k \leqslant j-1) \\
&\qquad + 2w_k \gamma_k^{-2} \sum_{i=1}^{j-1} I(k \leqslant i-1) h_i^{-1} \log \gamma_i \Bigg\}
\end{aligned}$$

and

$$-\frac{\partial^2 l}{\partial \phi_k \partial \phi_l} = -\sum_{j=1}^{m} I_j \left\{ \frac{\partial^2 \log p_j}{\partial \phi_k \partial \phi_l} - \frac{\partial^2 R_j}{\partial \phi_k \partial \phi_l} \right\},$$

where, for $k > l$,

$$\begin{aligned}
\frac{\partial^2 R_j}{\partial \phi_k \partial \phi_l} &= \sigma_1^{-1} \Bigg\{ h_j^{-1} w_l \gamma_l^{-1} \left[\phi_j^{-1} \log T_j - S_j T_j^{-1} \right] I(k=j) \\
&\qquad + w_k \gamma_k^{-1} w_l \gamma_l^{-1} I(k \leqslant j-1) h_j^{-1} \log T_j \Bigg\}, \\
\frac{\partial^2 \log p_j}{\partial \phi_k \partial \phi_l} &= -\sigma_1^{-1} \Bigg\{ h_k^{-1} w_l \gamma_l^{-1} \left[\phi_k^{-1} \log \gamma_k - w_k \gamma_k^{-1} \right] I(k \leqslant j-1) \\
&\qquad + w_k \gamma_k^{-1} w_l \gamma_l^{-1} \sum_{i=1}^{j-1} I(k \leqslant i-1) h_i^{-1} \log \gamma_i \Bigg\}.
\end{aligned}$$

Only terms containing S_j and/or T_j involve Y, so element (k,l) of the expected information matrix is of the form

$$\begin{aligned}
-\mathrm{E}\left(\frac{\partial^2 l}{\partial \theta_k \partial \theta_l} \right) &= \sum_{j=1}^{m} \int_{v_j}^{v_{j+1}} \left[A_j^{kl} + B_j^{kl}(y) \right] p_j f_j(y) \, \mathrm{d}y, \\
&= \sum_{j=1}^{m} A_j^{kl} q_j + \sum_{j=1}^{m} \int_{v_j}^{v_{j+1}} B_j^{kl}(y) \, p_j f_j(y) \, \mathrm{d}y,
\end{aligned}$$

for A_j^{kl} and B_j^{kl} implied by the equations above and where $q_j = P(v_j < Y < v_{j+1}) = p_j\left(1-\gamma_j^{-1/\xi_j}\right)$. Depending on k and l the second term involves some or all of the integrals

$$\int_{v_j}^{v_{j+1}} s_j t_j^{-1} f_j(y)\,\mathrm{d}y = \sigma_j(1+\xi_j)^{-1} + I(j<m)\phi_j^{-1}\left\{\gamma^{-(1+1/\xi_j)}(1+\xi_j)^{-1} - \gamma_j^{1/\xi_j}\right\},$$

$$\int_{v_j}^{v_{j+1}} s_j^2 t_j^{-2} f_j(y)\,\mathrm{d}y, = 2\sigma_j^2(1+\xi_j)^{-1}(1+2\xi_j)^{-1} - I(j<m)\phi_j^{-2} \times\left\{\gamma_j^{-1/\xi_j} - 2\gamma_j^{-(1+1/\xi_j)}(1+\xi_j)^{-1} + \gamma^{-(2+1/\xi_j)}(1+2\xi_j)^{-1}\right\},$$

$$\int_{v_j}^{v_{j+1}} \log t_j f_j(y)\,\mathrm{d}y = \xi_j - I(j<m)\gamma_j^{-1/\xi_j}\left\{\xi_j + \log\gamma_j\right\}.$$

These expressions have been checked using numerical differentiation and integration. For a random sample $(y_1,\ldots,y_n)$ from density (4) the score function is $\sum_{i=1}^n u_i$ and the expected information is $n\,i_1$.

References

Anderson CW (1990) Discussion of 'Models for exceedances over high thresholds (with discussion)' by A. C. Davison and R. L. Smith. J R Statist Soc B52:425–426

Coles SG (2001) An Introduction to statistical modelling of extreme values. Springer, London

Davison AC, Smith RL (1990) Models for exceedances over high thresholds (with discussion). J R Statist Soc B52:393–442

Fawcett L, Walshaw D (2012) Estimating return levels from serially dependent extremes. Environmetrics 23:272–283

Freedman DA (2007) How can the score test be inconsistent? The American Statistician 61:291–295

Morgan BJT, Palmer KJ, Ridout MS (2007) Negative score test statistic. The American Statistician 61:285–288

Northrop PJ, Jonathan P (2011) Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights. Environmetrics 22(7):799–809

Oceanweather Inc (2005) GOMOS – USA Gulf of Mexico Oceanographic Study, Northern Gulf of Mexico Archive

Pickands J (1971) The two-dimensional Poisson process and extremal processes. J App Prob 8:745–756

Pickands J (1975) Statistical inference using extreme order statistics. Ann Stat 3:119–131

Scarrott C, MacDonald A (2012) A review of extreme value threshold estimation and uncertainty quantification. REVSTAT - Statistical Journal 10(1):33–60

Smith RL (1985) Maximum likelihood estimation in a class of non-regular cases. Biometrika 72:67–92

Süveges M, Davison AC (2010) Model misspecification in peaks over threshold analysis. Ann Appl Statist 4:203–221

Wadsworth JL, Tawn JA (2012) Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling. J R Statist Soc B74(3):543–567