Submitted to the Annals of Applied Statistics

AUTOMATED THRESHOLD SELECTION FOR EXTREME VALUE ANALYSIS VIA ORDERED GOODNESS-OF-FIT TESTS WITH ADJUSTMENT FOR FALSE DISCOVERY RATE

By Brian Bader[*,†], Jun Yan[†,§], and Xuebin Zhang[‡]

KPMG LLP[*], University of Connecticut[†], and Environment and Climate Change Canada[‡]

Threshold selection is a critical issue for extreme value analysis with threshold-based approaches. Under suitable conditions, exceedances over a high threshold have been shown to follow the generalized Pareto distribution (GPD) asymptotically. In practice, however, the threshold must be chosen. If the chosen threshold is too low, the GPD approximation may not hold and bias can occur. If the threshold is chosen too high, reduced sample size increases the variance of parameter estimates. To process batch analyses, commonly used selection methods such as graphical diagnostics are subjective and cannot be automated. We develop an efficient technique to evaluate and apply the Anderson–Darling test to the sample of exceedances above a fixed threshold. In order to automate threshold selection, this test is used in conjunction with a recently developed stopping rule that controls the false discovery rate in ordered hypothesis testing. Previous attempts in this setting do not account for the issue of ordered multiple testing. The performance of the method is assessed in a large scale simulation study that mimics practical return level estimation. This procedure was repeated at hundreds of sites in the western US to generate return level maps of extreme precipitation.

1. Introduction. Extreme value analysis has wide applications in a variety of fields, such as hydrology (e.g., Katz, Parlange and Naveau, 2002) and climatology (e.g., Davison and Smith, 1990; Kharin et al., 2013), and dates back to Fisher and Tippett (1928). A major goal of inferences in these fields is to estimate the probability of extreme events, often expressed in terms of return level and return period. A return level with a return period of $T = 1/p$ years is a high value that is exceeded with probability p. In other words, the average number of events exceeding this level within a T-year period is one. Commonly used in extreme value analysis, threshold-based methods involve modeling data exceeding a suitably chosen high threshold with the generalized Pareto distribution (GPD) (Balkema and De Haan, 1974; Pickands, 1975). Choice of the threshold is critical in obtaining accurate estimates of model parameters and return levels. The threshold should

[§]Research partially supported by NSF grant DMS 1521730, University of Connecticut Research Excellence Program, and Environment Canada. The authors thank Prof. Vartan Choulakian for the discussion and insight on approximating the tails of the null distribution of the Anderson–Darling and Cramér–von Mises statistics for generalized Pareto distributed data.

Keywords and phrases: batch analysis, exceedance diagnostic, specification test, stopping rule

be chosen high enough for the excesses to be well approximated by the GPD to minimize bias, but not so high to substantially increase the variance of the estimator due to reduction in the sample size (the number of exceedances).

Although it is widely accepted in the statistics community that the peaks-over-threshold (POT) approach uses data more efficiently than the block maxima method (e.g., Caires, 2009; Wang, 1991), it is less utilized in some fields such as climatology. Even when appropriate data is available for use with the POT approach, there is a lack of efficient procedures that can be applied consistently and automatically to select the threshold in each sample, sometimes numbering in the hundreds or thousands (e.g., Kharin et al., 2007, 2013). As a motivating application, consider mapping the annual return levels of daily precipitation for the three west coastal US states of California, Oregon, and Washington. For the three states, one needs to repeat the estimation procedure, including threshold selection, at each of the hundreds of stations. For the whole US, thousands of sites would need to be processed. A graphical based diagnosis will be difficult to apply consistently across many sites and is clearly impractical. It is desirable to have an intuitive automated threshold selection procedure to use with POT analysis.

Many threshold selection methods are available in the literature; see Scarrott and MacDonald (2012), Caeiro and Gomes (2015) and Langousis et al. (2016) for recent reviews. Graphical diagnosis methods are commonly used (e.g., Davison and Smith, 1990; Drees, De Haan and Resnick, 2000; Coles, 2001; Scarrott and MacDonald, 2012), but can be quite subjective and not appropriate as an automated procedure. Recently, Northrop, Attalides and Jonathan (2017) take a unique approach by applying Bayesian model averaging to combine inferences from multiple thresholds in order to reduce the sensitivity in estimates from using a single, fixed threshold. Other selection methods can be grouped into various categories. One is based on the asymptotic results about estimators of properties of the tail distribution. Langousis et al. (2016) detail a few of these procedures for threshold selection, such as the Jackson (Jackson, 1967) and Lewis (Lewis, 1965) kernel statistics as modified in Goegebeur, Beirlant and de Wet (2008) (based on the Hill estimator) and an automated version of the mean residual life (MRL) plot. Computational, resampling-based estimators require sufficient computing resources, may involve tuning parameters (Danielsson et al., 2001), and in some cases is not satisfactory for small samples (Ferreira, de Haan and Peng, 2003).

A second category of methods are based on goodness-of-fit of the GPD, where the threshold is selected as the lowest level above which the GPD provides adequate fit to the exceedances (e.g., Davison and Smith, 1990;

Dupuis, 1999; Choulakian and Stephens, 2001; Northrop and Coleman, 2014; Langousis et al., 2016). Goodness-of-fit tests are simple to understand and perform, but error control, however, is challenging because of the ordered nature of the hypotheses, and the usual methods from multiple testing such as false discovery rate (FDR) (e.g., Benjamini, 2010a,b) cannot be directly applied. This has not been addressed to the best of our knowledge. Methods in the third category are based on mixtures of a GPD for the tail and another distribution for the "bulk" joined at the threshold (e.g., MacDonald et al., 2011; Wadsworth and Tawn, 2012; Naveau et al., 2016). Treating the threshold as a parameter to estimate, these methods can account for the uncertainty from threshold selection in inferences. However, care is needed to ensure that the bulk and tail models are robust to one another in the case of misspecification.

The simple naive method is *a priori* or fixed threshold selection based on expertise on the subject matter at hand. Various rules of thumb have been suggested; for example, selecting the top 10% of the data (e.g., DuMouchel, 1983), or the top square root of the sample size (e.g., Ferreira, de Haan and Peng, 2003). Such one rule for all is not ideal in climate applications where high heterogeneity in data properties is the norm. The proportion of the number of rain days can be very different from wet tropical climates to dry subtropical climates; therefore the number of exceedances over the same time period can be very different across different climates. Additionally, the probability distribution of daily precipitation can also be different in different climates, affecting the speed of the tail convergence to the GPD (Raoult and Worms, 2003).

We propose an automated threshold selection procedure based on a sequence of goodness-of-fit tests with error control for ordered, multiple testing. The recently developed stopping rules in G'Sell et al. (2015), which control the FDR (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001) for ordered hypotheses, are adapted for use in this setting. They are applied to the Anderson–Darling (AD) goodness-of-fit test at each candidate threshold sequentially from low to high. The application is challenging in that the asymptotic null distribution of the testing statistic is unwieldy (Choulakian and Stephens, 2001), and that parametric bootstrap puts bounds on the approximate p-values which can be shown to reduce power of the stopping rules. We propose a fast approximation of the p-value based on the results of Choulakian and Stephens (2001) to facilitate the application. The performance of the procedures are investigated in a large scale simulation study, and recommendations are made. The procedure is applied to precipitation return level mapping of three west coastal states

of the US. Interesting findings are revealed from different stopping rules. The automated threshold selection procedure has applications in various fields, especially when a consistent application for batch processing of massive datasets and of many sites is needed.

The outline of the paper is as follows. Section 2 presents the generalized Pareto model, its theoretical justification, and how to apply the automated sequential threshold testing procedure. Section 3 introduces the tests proposed to be used in the automated testing procedure. A simulation study demonstrates the power of the tests for a fixed threshold under various misspecification settings and it is found that the AD test is most powerful in the vast majority of cases. A large scale simulation study in Section 4 demonstrates performance of the stopping rules for multiple ordered hypotheses, under a plausible misspecified distribution and is compared with competing methods. In Section 5, we return to our motivating application and derive return levels for extreme precipitation at hundreds of west coastal US stations to demonstrate the usage of our method and some practical considerations. A final discussion is delivered in Section 6.

2. Automated Sequential Testing Procedure. Threshold methods for extreme value analysis are based on that, under general regularity conditions, the only possible non-degenerate limiting distribution of properly rescaled exceedances of a threshold u is the GPD as $u \to \infty$ (e.g., Pickands, 1975). The GPD(σ_u, ξ) has cumulative distribution function

$$F(y|\theta) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\sigma_u}\right)^{-1/\xi} & \xi \neq 0, \quad y > 0, \quad 1 + \frac{\xi y}{\sigma_u} > 0, \\ 1 - \exp\left(-\frac{y}{\sigma_u}\right) & \xi = 0, \quad y > 0, \end{cases} \tag{1}$$

where $\theta = (\sigma_u, \xi)$, ξ is a shape parameter, and $\sigma_u > 0$ is a threshold-dependent scale parameter. The GPD also has the property that for some threshold $v > u$, the excesses follow a GPD with the same shape parameter, but a modified scale $\sigma_v = \sigma_u + \xi(v - u)$. For the remainder of the text, the subscript u associated with σ is dropped for ease of presentation.

Let $X_1, \ldots, X_n$ be a random sample of size n. If u is sufficiently high, the exceedances $Y_i = X_i - u$ for all i such that $X_i > u$ are approximately a random sample from a GPD. The question is to find the lowest threshold such that the GPD fits the sample of exceedances over this threshold adequately. Our solution is to combine sequential goodness-of-fit testing with adjustments for multiplicity in the ordered setting. The first component is similar to the approach taken in Choulakian and Stephens (2001) or model specification testing (e.g., Northrop and Coleman, 2014; Wadsworth, 2016)

for the GPD to the exceedances over each candidate threshold in an increasing order. The second component, multiple testing in the special ordered setting, is handled by the stopping rules in G'Sell et al. (2015).

Consider a fixed set of candidate thresholds $u_1 < \ldots < u_l$. For each threshold, there will be n_i excesses, $i = 1, \ldots, l$. The sequence of null hypotheses can be stated as

$H_0^{(i)}$: The distribution of the n_i exceedances above u_i follows the GPD.

For a fixed u_i, many tests are available for this $H_0^{(i)}$. An automated procedure can begin with u_1 and continue until some threshold u_i provides an acceptance of $H_0^{(i)}$ (Choulakian and Stephens, 2001; Thompson et al., 2009). The problem, however, is that unless the test has high power, an acceptance may happen at a low threshold by chance and, thus, the data above the chosen threshold is contaminated. One could also begin at the threshold u_l and descend until a rejection occurs, but this would result in an increased type I error rate. The multiple testing problem obviously needs to be addressed, and the issue here is especially challenging because these tests are ordered; if $H_0^{(i)}$ is rejected, then $H_0^{(k)}$ has been rejected for all $1 \leq k < i$. Despite the extensive literature on multiple testing and the more recent developments on FDR control and its variants (e.g., Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Benjamini, 2010a,b), no definitive procedure has been available for error control in ordered tests until the recent work of G'Sell et al. (2015).

We adapt the ForwardStop rule of G'Sell et al. (2015) to the sequential testing of (ordered) null hypotheses $H_1, \ldots, H_l$. Let $p_1, \ldots, p_l \in [0, 1]$ be the corresponding p-values of the l hypotheses. G'Sell et al. (2015) transform the sequence of p-values to a monotone sequence and then apply the original method of Benjamini and Hochberg (1995) on the monotone sequence. The rejection rule is constructed by returning a cutoff $\hat{k}$ such that $H_1, \ldots, H_{\hat{k}}$ are rejected. If no $\hat{k} \in \{1, \ldots, l\}$ exists, then no rejection is made. ForwardStop is given by

$$\hat{k}_{\mathrm{F}} = \max \left\{ k \in \{1, \ldots, l\} : -\frac{1}{k} \sum_{i=1}^{k} \log(1 - p_i) \leq \alpha \right\}, \tag{2}$$

where α is a pre-specified level.

Under the assumption of independence among the tests, ForwardStop was shown to control the FDR at level α. In our setting, stopping at k implies that goodness-of-fit of the GPD to the exceedances at the first k thresholds $\{u_1, \ldots, u_k\}$ is rejected. In other words, the first set of k null hypotheses

$\{H_1, \ldots, H_k\}$ is rejected. At each $H_0^{(i)}$, ForwardStop is a transformed average of the previous and current p-values.

Another stopping rule developed in G'Sell et al. (2015), StrongStop, could be applied in the same manner. However, it provides stronger error control than ForwardStop; it controls the familywise error rate instead of FDR. In that regard, StrongStop is less desirable for this application as the stricter error control leads to decreased power-to-reject and generally results in selected thresholds that are too low.

ForwardStop, combined with the sequential hypothesis testing, provides an automated selection procedure — all that is needed is the level of desired error control and a set of thresholds. A caveat is that the p-values of the sequential tests here are dependent, unlike the setup of G'Sell et al. (2015). Nonetheless, ForwardStop may still provide some reasonable error control as its counter part in the non-sequential multiple testing scenario (Benjamini and Yekutieli, 2001; Blanchard and Roquain, 2009). A simulation study is carried out in Section 4 to assess the empirical properties.

Additionally, it is of worth to note that there are two potential factors in the application to real data that may directly affect threshold selection using the ForwardStop procedure. First, by using critical values of goodness-of-fit tests intended for continuous data, applied to quantized data, resulting p-values have the potential to be underestimated (e.g., Langousis et al., 2016). Second, ignoring serial dependence in the excesses can lead to spurious precision in inferences about the GP parameters. Although ignoring such dependence does not introduce bias, it does underestimate standard errors (Fawcett and Walshaw, 2007), which in turn can influence the variability in goodness-of-fit test critical values.

3. The Tests.

The automated procedure can be applied with any valid test for each hypothesis $H_0^{(i)}$ corresponding to threshold u_i. Four existing goodness-of-fit tests that can be used are presented. Because the stopping rules are based on transformed p-values, it is desirable to have testing statistics whose p-values can be accurately measured; bootstrap based tests that put a lower bound on the p-values (1 divided by the bootstrap sample size) may lead to premature stopping. For the remainder of this section, the superscript i is dropped. We consider the goodness-of-fit of GPD to a sample of size n of exceedances $Y = X - u$ above a fixed threshold u.

3.1. *Anderson–Darling and Cramér–von Mises Tests.*

The AD and the Cramér–von Mises tests for the GPD have been studied in detail (Choulakian and Stephens, 2001). Let $\hat{\theta}_n$ be the maximum likelihood estimator (MLE) of θ under H_0 from the the observed exceedances. Make the probability

integral transformation based on $\hat{\theta}_n$ $z_{(i)} = F(y_{(i)}|\hat{\theta}_n)$, as in (1), for the order statistics of the exceedances $y_{(1)} < \ldots < y_{(n)}$. The AD statistic is

$$A_n^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\Big[\log(z_{(i)}) + \log(1 - z_{(n+1-i)})\Big].$$

The Cramér–von Mises statistic is

$$W_n^2 = \sum_{i=1}^{n}\big[z_{(i)} - \frac{2i-1}{2n}\big]^2 + \frac{1}{12n}.$$

The AD statistic is a slight modification of the Cramér–von Mises statistic, with more weight to observations in the tail of the distribution.

The asymptotic distributions of A_n^2 and W_n^2 are unwieldy, both being sum of weighted chi-squared variables with one degree of freedom with weight found from the eigenvalues of an integral equation (Choulakian and Stephens, 2001, Section 6). The distributions depend only on the estimate of ξ. The tests are often applied by referring to a table of a few upper tail percentiles of the asymptotic distributions (Choulakian and Stephens, 2001, Table 2), or through bootstrap. In either case, the p-values are truncated by a lower bound. Such truncation of a smaller p-value to a larger one can be proven to weaken the ForwardStop rule given in (2). In order to apply these tests in the automated sequential setting, more accurate p-values for the tests are needed.

We provide two remedies to the table in Choulakian and Stephens (2001). First, for ξ values in the range of $(-0.5, 1)$, which is applicable for most applications, we enlarge the table to a much finer resolution through a pre-set Monte Carlo method. For each ξ value from -0.5 to 1 incremented by 0.01, 2,000,000 replicates of A_n^2 and W_n^2 are generated with sample size $n = 1,000$ to approximate their asymptotic distributions. A grid of upper percentiles from 0.999 to 0.001 for each ξ value is produced and saved in a table for fast future reference. Therefore, if estimate $\hat{\xi}_n$ and the test statistic falls in the range of the table, the p-value is computed via interpolation on the logarithmic scale. That is, the p-value for an observed test statistic is found by taking the distance (on the log-scale) between the two nearest critical values found in the table, and scaling the distances to a maximum range of 0.001. For example, if the test statistic falls at the one-quarter mark between two critical values (on the log-scale), its p-value is equal to the smaller critical value's p-value + 0.00025.

The second remedy is for observed test statistics that are greater than that found in the table (implied p-value less than 0.001). As Choulakian pointed

out (in a personal communication), the tails of the asymptotic distributions are exponential, which can be confirmed using the available tail values in the table. For a given $\hat{\xi}_n$, regressing $-\log(\text{p-value})$ on the upper tail percentiles in the table, for example, from 0.05 to 0.001, gives a linear model that can be extrapolated to approximate the p-value of observed statistics outside of the range of the table. This approximation of extremely small p-values help reduce loss of power in the stopping rules.

The two remedies make the two tests very fast and are applicable for most applications with $\xi \in (-0.5, 1)$. For ξ values outside of $(-0.5, 1)$, although slow, one can use parametric bootstrap to obtain the distribution of the test statistic, understanding that the p-value has a lower bound. The methods are implemented in R package `eva` (Bader and Yan, 2015).

3.2. *Moran's Test and Rao's Score Test.* Two other tests are compared for reference. Moran's goodness-of-fit test is a byproduct of the maximum product spacing (MPS) estimation for estimating the GPD parameters. The maximized objective function (i.e., evaluated at the MPS estimators) is Moran's statistic (Moran, 1953). Cheng and Stephens (1989) showed that, under the GPD null hypothesis, Moran's statistic when properly centered and scaled, has an asymptotic chi-square approximation. Wong and Li (2006) show empirically the test holds its size for samples as small as ten.

Northrop and Coleman (2014) considered a piecewise (varying) representation of the shape parameter of the GPD at intervals between a set of thresholds. Under H_0 it is assumed that the shape parameter holds a constant value across all intervals. Tests of this hypothesis are developed based on a multiple-threshold penultimate model, using a construction of Rao's score test.

3.3. *A Power Study.* The power of the four goodness-of-fit tests were examined in an individual, non-sequential testing framework. The data generating schemes in Choulakian and Stephens (2001) were used, some of which are very difficult to distinguish from the GPD:

- Gamma with shape 2 and scale 1.
- Standard lognormal distribution (mean 0 and scale 1 on log scale).
- Weibull with scale 1 and shape 0.75.
- Weibull with scale 1 and shape 1.25.
- 50/50 mixture of $\text{GPD}(1, -0.4)$ and $\text{GPD}(1, 0.4)$.
- 50/50 mixture of $\text{GPD}(1, 0)$ and $\text{GPD}(1, 0.4)$.
- 50/50 mixture of $\text{GPD}(1, -0.25)$ and $\text{GPD}(1, 0.25)$.

Finally, the $\text{GPD}(1, 0.25)$ was also used to check the type I error rate. Four

sample sizes were considered: 50, 100, 200, 400. For each scenario, 10,000 samples are generated. The four tests were applied to each sample, with a rejection recorded if the p-value is below 0.05. An additional requirement of the score test is to select a set of intervals to constitute the piecewise representation of the shape parameter (Section 1.2, Northrop and Coleman, 2014); we set these according to the deciles of the generated data. The likelihood under each specification was maximized with the default setting of R function `optim`, the Nelder–Mead method.

[Table 1 about here.]

The rejection rates are summarized in Table 1. Samples in which the MLE failed were removed, which accounts for roughly 10.8% of the Weibull samples with shape 1.25 and sample size 400, and around 10.7% for the Gamma distribution with sample size 400. Decreasing the sample size in these cases actually decreases the percentage of failed MLE samples. This may be due to the shape of these two distributions, which progressively become more distinct from the GPD as their shape parameters increase. In particular, the Weibull changes shape dramatically as it crosses 1, which is the case here (0.75 to 1.25). In the other distribution cases, no setting resulted in more than a 0.3% failure rate. As expected, all tests appear to hold their sizes, and their powers all increase with sample size. The mixture of two GPDs is the hardest to detect. For the GPD mixture of shape parameters 0 and 0.4, quantile matching between a single large sample of generated data and the fitted GP distribution shows a high degree of similarity. In the vast majority of cases, the AD test appears to have the highest power, followed by the Cramér–von Mises test. The R-language code used to perform this power study can be found in the Supplementary Materials (Bader, Yan and Zhang, 2017).

4. Simulation Study of the Automated Procedure. It is of interest to investigate the performance of ForwardStop versus several competing methods under misspecification. Empirical work on rainfall extremes (e.g., Papalexiou and Koutsoyiannis, 2013; Serinaldi and Kilsby, 2014) finds that the right tail for daily rainfall is better described by heavy tail distributions than exponential types. Following the study of Roth, Jongbloed and Buishand (2016), data is generated from a distribution characterized by a hazard function which transits smoothly from the hazard function of a Weibull distribution to that of a GPD, with the transition occurring near some threshold u. Let $h_1(x) = \kappa\beta^{-\kappa}x^{\kappa-1}$, $x > 0$, be the hazard function of a Weibull distribution with parameters (κ, β). Let $h_2(x) = (\sigma + \xi(x-u))^{-1}$, $x \geq u$ be the hazard function of a GPD with parameters (ξ, σ). A smooth

transition from $h_1(\cdot)$ to $h_2(\cdot)$, with a transition period of length between $(u, u+\epsilon)$, defines a new hazard function

$$h(x) := h_1(x)\eta\Big(\frac{x-u}{\epsilon}\Big) + h_2(x)\Big(1-\eta\Big(\frac{x-u}{\epsilon}\Big)\Big),$$

where

$$\eta(x) = \begin{cases} 1, & x \le 0, \\ 2x^3 - 3x^2 + 1, & 0 < x < 1, \\ 0, & x \ge 1, \end{cases}$$

to ensure that h has a continuous derivative. The distribution function of the data generating mechanism is then

$$F(x) = 1 - \exp\big(-H(x)\big) = 1 - \exp\Big(\int_0^x h(z)\mathrm{d}z\Big).$$

More detailed properties of this distribution are discussed in Roth, Jongbloed and Buishand (2016, Section 5.1); a similar construction consisting of a GPD mixture distribution was proposed by Holden and Haug (2009). Four different thresholds are considered — $(u_1, u_2, u_3, u_4) = (17.70, 11.00, 5.91, 3.64)$ with $(\sigma_{u_1}, \sigma_{u_2}, \sigma_{u_3}, \sigma_{u_4}) = (11.51, 10.45, 6.80, 4.55)$. All four distributions share the same transition period length $\epsilon = 0.5$, GPD parameter $\xi = 0.15$ and Weibull parameters $(\kappa, \beta) = (0.45, 1)$. Starting from $u+\epsilon$, the tail of the distribution is completely the tail of the GPD. Our selections correspond to the 97.5th, 95th, 90th, and 85th percentiles, respectively. The distribution and density functions can be seen in Figure 1. For each setting, $B = 1,500$ samples were generated from this distribution with $n = 4,600$, which can be thought of as 50 years of seasonal daily observations (92 days each season).

[Fig 1 about here.]

The main quantity of interest is the N-year return level, defined for the GPD (e.g., Coles, 2001, Section 4.3.3) as

$$z_N = \begin{cases} u + \frac{\sigma}{\xi}[(Nn_y\zeta_u)^\xi - 1], & \xi \neq 0, \\ u + \sigma\log(Nn_y\zeta_u), & \xi = 0, \end{cases} \tag{3}$$

for a given threshold u, where n_y is the number of observations per year, and ζ_u is the rate, or proportion of the data exceeding u. In precise terms, this is a high quantile of the GPD, which has the interpretation as the level expected to be exceeded in a single year with probability $1/N$. In estimation, ζ_u is set to the proportion of threshold exceedances.

In addition to ForwardStop, three competing methods in return level estimation were used for comparison. The first was the so-called 'rules-of-thumb' (e.g., Ferreira, de Haan and Peng, 2003; DuMouchel, 1983), which simply uses a fixed (upper) fraction of the data to fit the GPD. In this setting, the top 15%, 10%, 5%, and square root (1.5%) of the sample size were used in each simulation to fit the GPD. The second, an automated procedure based on the MRL plot using weighted least squares (WLS) and detailed in Langousis et al. (2016, Section 2.2), chooses the threshold that minimizes the MSE of the WLS fit. The third was the unadjusted alternative to ForwardStop. It can be implemented two ways — both relying on the raw p-values from each threshold tested and proceeding sequentially. Specifically, the first version (denoted as Raw Up) begins at the lowest threshold and selects the first (lowest) threshold which is accepted. If all are rejected, the maximum threshold is selected. The second (denoted as Raw Down) starts from the largest threshold and proceeds until a rejection of the test occurs; then the threshold before the rejection is selected. If a rejection occurs on the first (highest) threshold, that threshold is used. ForwardStop has the same direction of operation as Raw Up, so, it also chooses the largest threshold if all are rejected.

The estimated 50-year return levels were compared using the chosen threshold. For each sample, two sets of percentiles were used to generate thresholds to test. The first takes 10 percentiles, in increments of 5 beginning at 50 and ending at 95, the second takes 20 percentiles, in increments of 2.5 beginning at 50 and ending at 97.5. Three significance levels $\alpha \in \{0.05, 0.10, 0.20\}$ were used for testing with ForwardStop and the two unadjusted procedures. For the automated MRL plot method, a threshold was selected for each set of thresholds. For a given return period N, the root mean squared error (RMSE) of an estimator is calculated as

$$\sqrt{\frac{1}{B}\sum_{i=1}^{B}(\hat{z}_N^i - z_N)^2}.$$

Figure 2 shows the RMSE comparison of all the methods considered; the corresponding plot for bias and variance (standard deviation) are available in the Supplementary Materials (Bader, Yan and Zhang, 2017).

[Fig 2 about here.]

As expected, for the 'top' methods with a fixed threshold, the resulting error in estimation is highly dependent on the chosen fixed threshold. The correct fixed threshold fixed method always has smallest RMSE. However,

when the threshold is fixed at an incorrect percentile, it is outperformed by ForwardStop and the unadjusted Raw Up/Down procedures. The MRL method has varying performance, dependent on the exceedance probability. Another property of the MRL is that a threshold is always selected and thus the steps for further diagnostic testing is unclear. The absolute bias for ForwardStop is generally lower than Raw Up/Down. Compared with ForwardStop, Raw Down performs better for small exceedance probabilities and worse for larger values (variance increases); Raw Up performs better for larger exceedance probabilities and worse for smaller values (bias increases). ForwardStop hedges against both cases and provides similar or smaller error than the better of Raw Up and Raw Down. This is an intuitive result — as the exceedance probability decreases, it is more likely for Raw Up to stop too early (increased bias). Similarly, as the exceedance probability increases, it is more likely for Raw Down to stop too high (increased variance). ForwardStop guards against either of these cases.

An alternative simulation study is reported in the Supplementary Materials (Bader, Yan and Zhang, 2017).

5. Return Level Mapping of Extreme Precipitation. Hydrologists are often interested in estimating return levels of extreme precipitation (Katz, Parlange and Naveau, 2002). These return levels are provided as maps for infrastructure design for risk management and land planning (Blanchet and Lehning, 2010; Lateltin and Bonnard, 1999). Estimating return levels across many sites consistently is often required to produce such maps. Our methodology uses automated threshold selection to generate return level maps through batch processing of data from a large number of sites. Here we consider a return level map of extreme precipitation in the three western US coastal states of California, Oregon, and Washington with diverse climates to demonstrate its application and practical usage. The automated procedure in Section 2 provides a quick and consistent way to obtain an accurate map without the need to inspect every site, and without taking a homogeneous approach.

Daily precipitation data is available for tens of thousands of surface sites around the world via the Global Historical Climatology Network (GHCN). A description of the data network can be found in Menne et al. (2012). After an initial screening to remove sites with less than 50 years of available data, there were 720 remaining sites across the three chosen coastal states. As the annual maximum daily amount of precipitation mainly occurs in winter, only the winter season (November to March) observations were used in modeling.

A major issue in the analysis of US precipitation data is the sensitivity of

statistical procedures in response to quantization. The United States precipitation data has a unit in millimeters (mm) and the data are recorded to the nearest hundredth of an inch. As a result, a rough rounding occurs with 0.2 and 0.3 mm intervals. The inflation effect of quantization on goodness-of-fit tests for the GPD (in particular, AD) has been well-known (e.g., Deidda and Puliga, 2006, 2009; Langousis et al., 2016). Quantization pushes the null distribution of the AD statistic to the right; the p-value obtained by positioning the observed statistic with quantized data to the null distribution from continuous data is smaller than it should be. Rougher rounding means bigger change in magnitude of the null distribution. Given a rounding level, the smaller the scale parameter of the null GPD, the bigger the increase in the magnitude of the AD test statistic. To accommodate the quantization issue, we seek to adjust the test statistic instead of the sampling distribution, because the sampling distribution depends on the scale parameters and needs to be approximated by Monte Carlo.

We approached the quantization issue by treating the observed data as uniformly interval censored. That is, for an observed value of $x_i > 0$, the actual observation is treated as interval censored by the interval $[x_i - \delta/2, x_i + \delta/2)$, where $\delta = 0.254$. To calculate the AD statistic, we perturbed each observed value x_i by a random draw from the uniform distribution over $[-\delta/2, \delta/2)$. We replicated the jittering process K times, and took the median of the resulting K AD statistics as our testing statistic. Its p-value was then found from the null distribution obtained for continuous data via the procedure in Section 3.1. As observed in the simulation study reported in the Supplementary Materials (Bader, Yan and Zhang, 2017), this jittering brings the AD statistic from quantized data much closer to the continuous version than the non-jittered version. The distribution of the jittered version of the AD statistic is approximated reasonably well by the distribution of the AD statistic for continuous data under varying degrees of quantization when σ_u is 'large'. Based on the experiments, sites are screened by taking $\sigma_{u=F^{-1}(0.7)} > 5.3$, which is the 25th percentile of estimated scale parameters estimated at all sites in an exploratory data analysis using the top 30% of the data at each site. This results in 501 sites to be used. A fully satisfactory solution to handle the quantization is noted as a topic of future work for its practical importance.

Another potential source of uncertainty in the goodness-of-fit testing procedure is the presence of serial dependence in the excesses. The goodness-of-fit tests assume independence in the excesses under the null hypothesis and thus significant departures may affect the testing conclusion. One way to check for this is the extremal index (Leadbetter et al., 1989), which is a

measure of the clustering of the underlying process at extreme levels and quantifies such dependence. It can take values from 0 to 1, with independent series exhibiting a value of exactly 1. To get a sense for the properties of series in this dataset, the extremal index was estimated using the so-called intervals estimator of Ferro and Segers (2003) for each of the 501 selected sites at the 70th percentile threshold via the R package `texmex` (Southworth and Heffernan, 2012). In summary, 199 sites (40%) have an estimated extremal index of 1, with a minimum value of 0.819. So the serial dependence was not considered largely influential.

Candidate thresholds were chosen based on the data percentiles. For each site, the set of thresholds was formed by taking the 70th to 98th percentiles in increments of 2, resulting in 15 thresholds to test. The 98th percentile is chosen as the upper limit to ensure a sufficient amount of data is available for parameter estimation. The FS rule is applied using significance level $\alpha = 0.05$. If all thresholds were rejected at a site, the 98th percentile was used to estimate the GPD parameters and corresponding return levels for that site. Out of the 501 sites, 33 (7%) had all thresholds rejected. This is an indication that these cases need to be looked at more thoroughly. One possible explanation is that the quantized data led to over rejection in the AD tests, and, hence, the threshold selected tends to be over high. If an indirect solution is desired, one may relax the distributional assumptions (e.g., Papastathopoulos and Tawn, 2013; Nadarajah and Eljabri, 2013) and follow the same procedure. It is of worth to compare the selected thresholds at each site found by various methods. Pairwise plots of selected thresholds for the FS, unadjusted, and MRL methods are presented in the Supplementary Materials (Bader, Yan and Zhang, 2017).

[Fig 3 about here.]

With the automatically selected thresholds for the 501 sites, the return levels were estimated. Figure 3 shows map of the 50-, 100-, and 250-year estimated return levels of the three states. Note that only a very small number of sites in eastern Washington/Oregon and the Great Valley of California show on the map, with low return levels. The screening process removed many sites from these specific geographic areas, which had a smaller percentage of precipitation days than the remaining 501 analyzed on average (27% versus 38%) in the winter season.. These areas are known dry areas because the moist air from the Pacific has already lost much of its moisture on the windward west side of the mountains, and as it descends on the leeward east side of the mountains, it warms adiabatically, making it less likely for the air to saturate and form precipitation. The selective feature of the

FS procedure even after the screening process is a desirable as it suggests not to fit GPD at even the highest threshold at these sites, a guard that is not available from those unconditionally applied, one-for-all rules.

6. Discussion. We propose an intuitive and comprehensive methodology for automated threshold selection in the peaks over threshold approach. In addition, it is less computationally intensive than some competing resampling or bootstrap procedures (Danielsson et al., 2001; Ferreira, de Haan and Peng, 2003). Automation and efficiency is required when prediction using the peaks-over-threshold approach is desired at a large number of sites. This is achieved through sequentially testing a set of thresholds for goodness-of-fit to the GPD. This general methodology has been applied previously (e.g., Choulakian and Stephens, 2001; Thompson et al., 2009); however these did not account for the multiplicity issue. That is, they selected a threshold by checking raw p-values against the desired significance level until a rejection occurred. We apply the recently developed stopping rule ForwardStop (G'Sell et al., 2015), which transforms the results of ordered, sequentially tested hypotheses to control the false discovery rate. There is a slight caveat in our setting, that the tests are not independent, but it can be demonstrated via simulation that the stopping rules in G'Sell et al. (2015) still provide reasonable error control here.

Four tests are compared in terms of power to detect departures from the GPD at a single, fixed threshold and it is found that the AD test has the most power in various non-null settings. Choulakian and Stephens (2001) derived the asymptotic null distribution of the AD test statistic. However this requires solving an integral equation. Our contribution, with some advice from Professor Choulakian, provides an approximate, but accurate and computationally efficient version of this test. To investigate the performance of ForwardStop in conjunction with the AD test, a large scale simulation study was conducted. Data is generated from a plausible distribution – misspecified below a certain threshold and generated from the null GPD above, smoothed via a transition function. In each replicate, the squared error is recorded for various parameters using ForwardStop and compared with competing alternative procedures. The results of this simulation are mixed – no procedure outperformed in all settings. ForwardStop is, however, less sensitive to the starting threshold than the unadjusted versions of sequential hypothesis testing (Raw Up and Down) and it is not forced to always select a threshold, unlike the MRL approach.

The entire methodology using ForwardStop is applied to daily precipitation data at hundreds of sites in three U.S. west coast states, with the

goal of creating a return level map. However, quantization present in the U.S. GHCN dataset introduces additional uncertainty and in this instance, increased conservatism when performing goodness-of-fit testing on the quantized data. To handle this, the data are treated as (uniformly) interval censored and an adjusted testing statistic is found via realized samples. As is discussed in the Supplementary Materials (Bader, Yan and Zhang, 2017), this solution is still conservative (i.e., smaller p-values, higher thresholds selected). The most precise way is to obtain the null distribution of the Anderson–Darling statistic under quantization, but this is cumbersome as the distribution depends on both the relative size of σ_u to the quantization level and ξ. Sites across eastern Washington/Oregon were screened out or estimated to have smaller return levels, which is consistent with the known climate of that region.

Temporal or covariate varying thresholds, as discussed in Roth et al. (2012) and Northrop and Jonathan (2011), is an obvious extension to this work. However, one particular complication that arises is performing model selection (i.e., what covariates to include), while concurrently testing for goodness-of-fit to various thresholds. It is clear that threshold selection will be dependent on the choice of model. Another possible extension involves testing for overall goodness-of-fit across sites (e.g., Roth, Jongbloed and Buishand, 2016). In this way, a fixed or quantile regression based threshold may be predetermined and then tested simultaneously across sites. In this setup both spatial and temporal dependence need to be taken into account. Handling this requires some care due to censoring (e.g., Dey and Yan, 2015, Section 2.5.2). In other words, it is not straightforward to capture the temporal dependence as exceedances across sites are not guaranteed to occur at the same points in time.

An important issue that needs to be addressed, both here and the field of extremes in general is quantization. It has been studied in detail (Deidda and Puliga, 2006, 2009; Langousis et al., 2016). The most obvious way to handle quantization is via simulation and/or bootstrap approaches. Direct application of this becomes cumbersome when testing the GPD for goodness-of-fit at a large number of thresholds since the testing statistic under quantization is dependent on both shape and scale parameters. Its effect has been reduced in this data analysis but future work is needed to directly accommodate quantization, particularly in the sequential testing setting.

SUPPLEMENTARY MATERIAL

Additional Simulation Results and Data Analysis
(doi: COMPLETED BY THE TYPESETTER; .pdf). Material consisting

of R code for the power study, additional simulation results, and analysis related to the application.

References.

BADER, B. and YAN, J. (2015). eva: Extreme Value Analysis with Goodness-of-Fit Testing R package version 0.1.2.

BADER, B., YAN, J. and ZHANG, X. (2017). Supplementary Materials to "Automated Threshold Selection for Extreme Value Analysis via Ordered Goodness-of-Fit Tests with Adjustment for False Discovery Rate". *Annals of Applied Statistics*.

BALKEMA, A. A. and DE HAAN, L. (1974). Residual Life Time at Great Age. *The Annals of Probability* **2** 792–804.

BENJAMINI, Y. (2010a). Discovering the False Discovery Rate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72** 405–416.

BENJAMINI, Y. (2010b). Simultaneous and Selective Inference: Current Successes and Future Challenges. *Biometrical Journal* **52** 708–721.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B* **57** 289–300.

BENJAMINI, Y. and YEKUTIELI, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics* **29** 1165–1188.

BLANCHARD, G. and ROQUAIN, É. (2009). Adaptive False Discovery Rate Control under Independence and Dependence. *The Journal of Machine Learning Research* **10** 2837–2871.

BLANCHET, J. and LEHNING, M. (2010). Mapping Snow Depth Return Levels: Smooth Spatial Modeling versus Station Interpolation. *Hydrology and Earth System Sciences* **14** 2527–2544.

CAEIRO, F. and GOMES, M. I. (2015). Threshold Selection in Extreme Value Analysis. In *Extreme Value Modeling and Risk Analysis: Methods and Applications* (D. K. Dey and J. Yan, eds.) 69–82. CRC Press.

CAIRES, S. (2009). A Comparative Simulation Study of the Annual Maxima and the Peaks-over-Threshold Methods Technical Report, SBW-Belastingen: subproject 'Statistics'. Deltares Report 1200264-002.

CHENG, R. C. H. and STEPHENS, M. A. (1989). A Goodness-of-Fit Test Using Moran's Statistic with Estimated Parameters. *Biometrika* **76** 385–392.

CHOULAKIAN, V. and STEPHENS, M. A. (2001). Goodness-of-Fit Tests for the Generalized Pareto Distribution. *Technometrics* **43** 478–484.

COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*, 1 ed. Springer.

DANIELSSON, J., DE HAAN, L., PENG, L. and DE VRIES, C. G. (2001). Using a Bootstrap Method to Choose the Sample Fraction in Tail Index Estimation. *Journal of Multivariate Analysis* **76** 226–248.

DAVISON, A. C. and SMITH, R. L. (1990). Models for Exceedances Over High Thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)* **52** 393–442.

DEIDDA, R. and PULIGA, M. (2006). Sensitivity of Goodness-of-Fit Statistics to Rainfall Data Rounding Off. *Physics and Chemistry of the Earth, Parts A/B/C* **31** 1240–1251.

DEIDDA, R. and PULIGA, M. (2009). Performances of Some Parameter Estimators of the Generalized Pareto Distribution over Rounded-Off Samples. *Physics and Chemistry of the Earth, Parts A/B/C* **34** 626–634.

Dey, D. K. and Yan, J., eds. (2015). *Extreme Value Modeling and Risk Analysis: Methods and Applications*. CRC Press.

Drees, H., De Haan, L. and Resnick, S. (2000). How to Make a Hill Plot. *The Annals of Statistics* **28** 254–274.

DuMouchel, W. H. (1983). Estimating the Stable Index α in Order to Measure Tail Thickness: A Critique. *The Annals of Statistics* **11** 1019–1031.

Dupuis, D. J. (1999). Exceedances over High Thresholds: A Guide to Threshold Selection. *Extremes* **1** 251–261.

Fawcett, L. and Walshaw, D. (2007). Improved Estimation for Temporally Clustered Extremes. *Environmetrics* **18** 173–188.

Ferreira, A., de Haan, L. and Peng, L. (2003). On Optimising the Estimation of High Quantiles of a Probability Distribution. *Statistics* **37** 401–434.

Ferro, C. A. and Segers, J. (2003). Inference for Clusters of Extreme Values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65** 545–556.

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

Goegebeur, Y., Beirlant, J. and de Wet, T. (2008). Linking Pareto-tail Kernel Goodness-of-Fit Statistics with Tail Index at Optimal Threshold and Second Order Estimation. *Revstat* **6** 51–69.

G'Sell, M. G., Wager, S., Chouldechova, A. and Tibshirani, R. (2015). Sequential Selection Procedures and False Discovery Rate Control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. Forthcoming.

Holden, L. and Haug, O. (2009). A Multidimensial Mixture Model for Unsupervised Tail Estimation NR-notat SAMBA/09/09. pp 29.

Jackson, O. (1967). An Analysis of Departures from the Exponential Distribution. *Journal of the Royal Statistical Society. Series B (Methodological)* 540–549.

Katz, R. W., Parlange, M. B. and Naveau, P. (2002). Statistics of Extremes in Hydrology. *Advances in Water Resources* **25** 1287–1304.

Kharin, V. V., Zwiers, F. W., Zhang, X. and Hegerl, G. C. (2007). Changes in Temperature and Precipitation Extremes in the IPCC Ensemble of Global Coupled Model Simulations. *Journal of Climate* **20** 1419–1444.

Kharin, V. V., Zwiers, F., Zhang, X. and Wehner, M. (2013). Changes in Temperature and Precipitation Extremes in the CMIP5 Ensemble. *Climatic Change* **119** 345–357.

Langousis, A., Mamalakis, A., Puliga, M. and Deidda, R. (2016). Threshold Detection for the Generalized Pareto Distribution: Review of Representative Methods and Application to the NOAA NCDC Daily Rainfall Database. *Water Resources Research* **52** 2659–2681.

Lateltin, O. and Bonnard, C. (1999). Hazard Assessment and Land-Use Planning in Switzerland for Snow Avalanches, Floods and Landslides Technical Report, World Meteorological Organization.

Leadbetter, M. R., Weissman, I., De Haan, L. and Rootzén, H. (1989). On Clustering of High Values in Statistically Stationary Series. *Proc. 4th Int. Meet. Statistical Climatology* **16** 217–222.

Lewis, P. A. (1965). Some results on tests for Poisson processes. *Biometrika* **52** 67–77.

MacDonald, A., Scarrott, C. J., Lee, D., Darlow, B., Reale, M. and Russell, G. (2011). A Flexible Extreme Value Mixture Model. *Computational Statistics & Data Analysis* **55** 2137–2157.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E. and Houston, T. G. (2012).

An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology* **29** 897–910.

Moran, P. A. P. (1953). The Random Division of an Interval-Part II. *Journal of the Royal Statistical Society. Series B (Methodological)* **15** 77–80.

Nadarajah, S. and Eljabri, S. (2013). The Kumaraswamy GP distribution. *Journal of Data Science* **11** 739–766.

Naveau, P., Huser, R., Ribereau, P. and Hannart, A. (2016). Modeling Jointly Low, Moderate, and Heavy Rainfall Intensities Without a Threshold Selection. *Water Resources Research* **52** 2753–2769.

Northrop, P. J., Attalides, N. and Jonathan, P. (2017). Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **66** 93–120.

Northrop, P. J. and Coleman, C. L. (2014). Improved Threshold Diagnostic Plots for Extreme Value Analyses. *Extremes* **17** 289–303.

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

Papalexiou, S. M. and Koutsoyiannis, D. (2013). Battle of extreme value distributions: A global survey on extreme daily rainfall. *Water Resources Research* **49** 187–201.

Papastathopoulos, I. and Tawn, J. A. (2013). Extended generalised Pareto models for tail estimation. *Journal of Statistical Planning and Inference* **143** 131–143.

Pickands, J. III (1975). Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics* **3** 119–131.

Raoult, J.-P. and Worms, R. (2003). Rate of Convergence for the Generalized Pareto Approximation of the Excesses. *Advances in Applied Probability* **35** 1007–1027.

Roth, M., Jongbloed, G. and Buishand, T. A. (2016). Threshold Selection for Regional Peaks-over-Threshold Data. *Journal of Applied Statistics* **43** 1291–1309.

Roth, M., Buishand, T. A., Jongbloed, G., Klein Tank, A. M. G. and van Zanten, J. H. (2012). A Regional Peaks-over-Threshold Model in a Nonstationary Climate. *Water Resources Research* **48**.

Scarrott, C. and MacDonald, A. (2012). A Review of Extreme Value Threshold Estimation and Uncertainty Quantification. *REVSTAT–Statistical Journal* **10** 33–60.

Serinaldi, F. and Kilsby, C. G. (2014). Rainfall extremes: Toward reconciliation after the battle of distributions. *Water resources research* **50** 336–352.

Southworth, H. and Heffernan, J. E. (2012). texmex: Threshold Exceedences and Multivariate Extremes R package version 1.3.

Thompson, P., Cai, Y., Reeve, D. and Stander, J. (2009). Automated Threshold Selection Methods for Extreme Wave Analysis. *Coastal Engineering* **56** 1013–1021.

Wadsworth, J. L. (2016). Exploiting Structure of Maximum Likelihood Estimators for Extreme Value Threshold Selection. *Technometrics* **58** 116–126.

Wadsworth, J. L. and Tawn, J. A. (2012). Likelihood-Based Procedures for Threshold Diagnostics and Uncertainty in Extreme Value Modelling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 543–567.

Wang, Q. J. (1991). The POT Model Described by the Generalized Pareto Distribution with Poisson Arrival Rate. *Journal of Hydrology* **129** 263–280.

Wong, T. S. T. and Li, W. K. (2006). A Note on the Estimation of Extreme Value Distributions Using Maximum Product of Spacings. In *Time Series and Related Topics* 272–283. Institute of Mathematical Statistics.

1 Brian Bader
KPMG LLP
560 Lexington Avenue
New York, NY 10022
E-mail: brianbader@kpmg.com

Jun Yan
University of Connecticut
Department of Statistics
215 Glenbrook Rd. U-4120
Storrs, CT 06269-4120
E-mail: jun.yan@uconn.edu

2 Xuebin Zhang
Environment and Climate Change Canada
Climate Research Division
4905 Dufferin Street
Downsview, Ontario M5H 5T4, Canada
E-mail: xuebin.zhang@canada.ca

TABLE 1

Empirical rejection rates of four goodness-of-fit tests for GPD under various data generation schemes with nominal size 0.05. GPDMix(a, b) refers to a 50/50 mixture of GPD(1, a) and GPD(1, b).

Sample Size	50				100			
Test	Score	Moran	AD	CVM	Score	Moran	AD	CVM
Gamma(2, 1)	7.6	9.7	47.4	43.5	8.0	14.3	64.7	59.7
LogNormal	6.0	5.9	13.3	8.6	5.4	8.2	28.3	23.4
Weibull(0.75)	11.5	7.8	55.1	23.5	12.1	9.5	65.1	39.4
Weibull(1.25)	6.6	11.3	29.1	27.3	5.6	12.5	20.8	19.2
GPDMix(−0.4, 0.4)	11.4	7.5	19.2	9.9	16.0	8.6	24.3	20.4
GPDMix(0, 0.4)	7.8	6.0	6.5	5.9	7.4	5.9	9.6	5.6
GPDMix(−0.25, 0.25)	8.1	6.5	6.0	7.4	8.9	6.5	11.1	8.4
GPD(1, 0.25)	6.9	5.5	6.7	6.1	5.0	5.2	5.2	5.2
Sample Size	200				400			
Test	Score	Moran	AD	CVM	Score	Moran	AD	CVM
Gamma(2, 1)	15.4	23.3	95.3	93.1	36.5	42.2	100.0	100.0
LogNormal	5.7	11.9	69.3	59.7	7.9	19.1	97.8	95.0
Weibull(0.75)	16.5	10.7	84.8	66.4	32.1	14.6	98.2	93.0
Weibull(1.25)	8.0	14.7	40.9	36.7	15.5	19.2	79.8	74.6
GPDMix(−0.4, 0.4)	31.5	9.7	45.1	44.0	63.8	11.9	79.9	80.2
GPDMix(0, 0.4)	8.9	6.5	8.8	7.3	12.3	6.2	10.8	10.3
GPDMix(−0.25, 0.25)	13.9	6.7	16.6	14.8	26.2	7.9	33.0	32.4
GPD(1, 0.25)	5.3	5.8	7.2	5.2	4.7	5.3	5.8	4.7

1

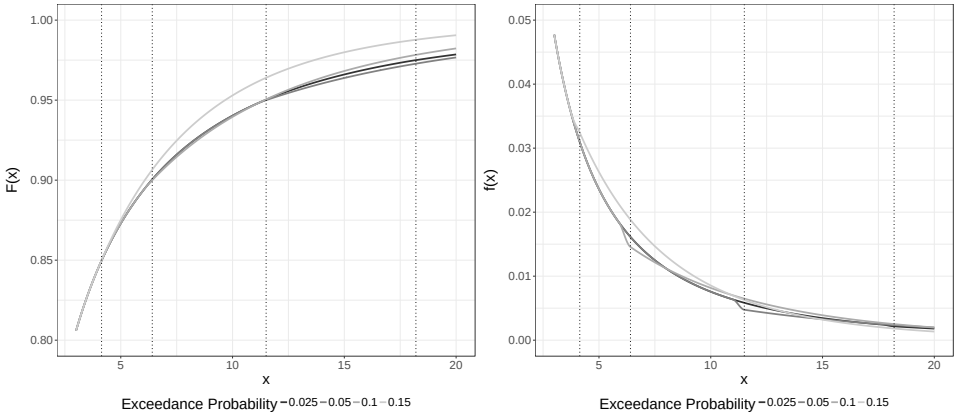

Fig 1: The cumulative and probability distributions used to generate data in the simulation study in Section 4. It is a Weibull distribution, with GPD tail. The start of the GP tail ($u + \epsilon$) is 18.20, 11.50, 6.41, and 4.14 for the 0.025, 0.05, 0.10, and 0.15 exceedance probabilities, respectively.

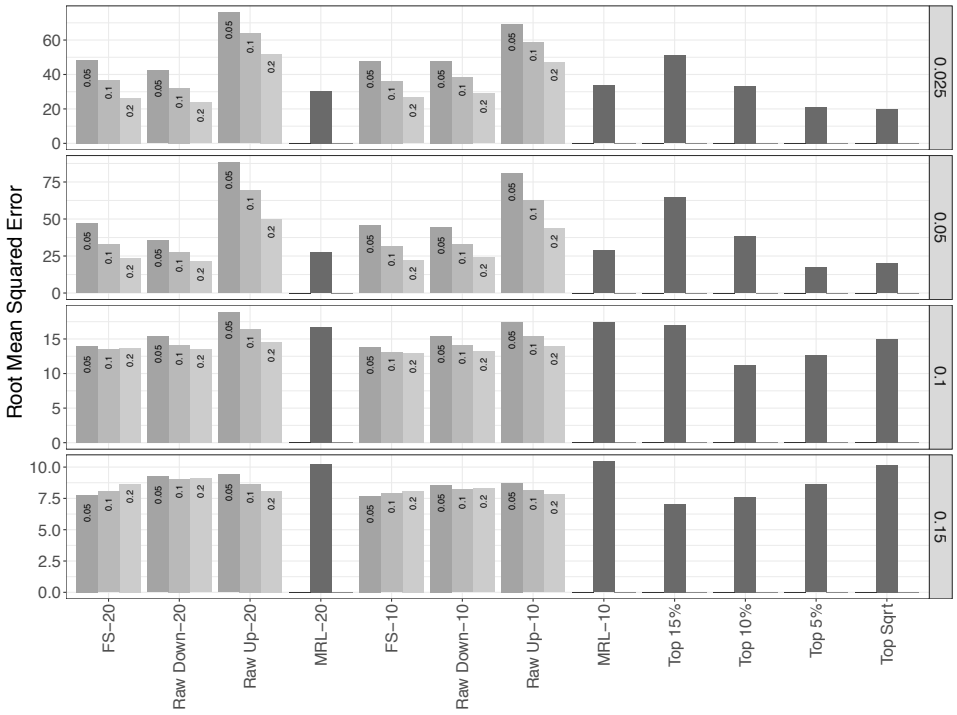

Fig 2: Root mean squared error (RMSE) of the estimators for the 50 year return level for the $B = 1,500$ simulations and each method described in Section 4. The attached -10 and -20 refer to the two sets of thresholds tested. The decimals on the right axis reference the exceedance probability. The decimals in the bars of Raw and FS (ForwardStop) correspond to the significance level α used in conjunction. Note that the 'top' methods are at a fixed threshold, thus no testing procedure is needed.

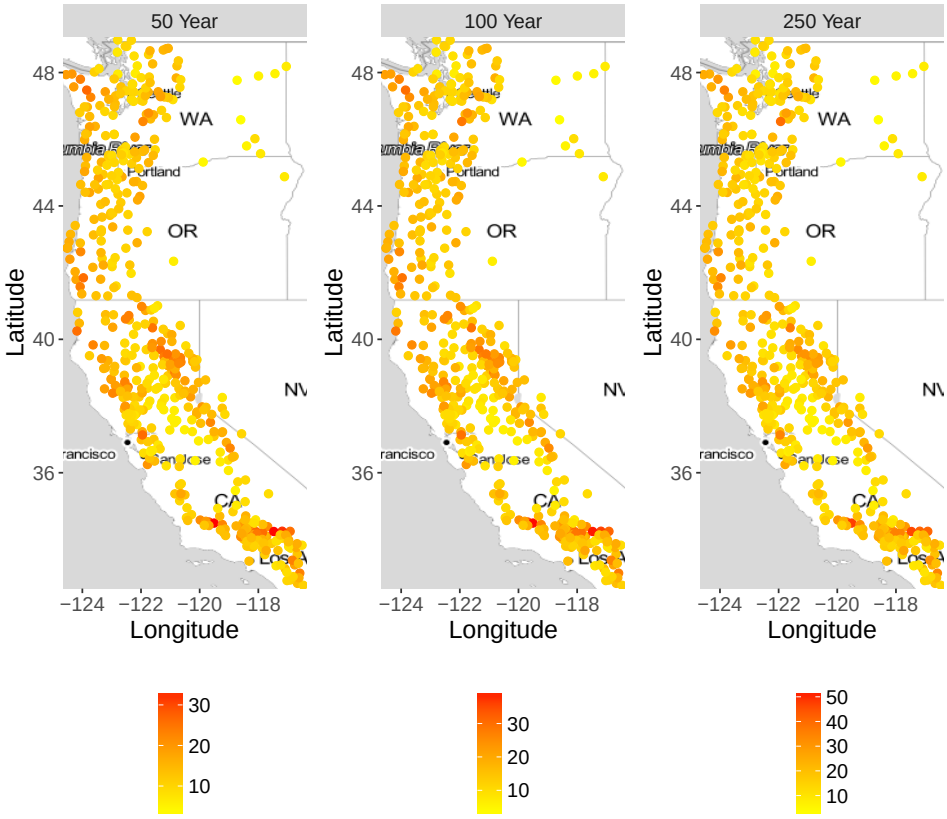

Fig 3: Estimated 50, 100, and 250 year daily precipitation return levels (cm) using the ForwardStop procedure and AD test, combined with jittering.