Extreme Value Modeling with Generalized Pareto Distributions for Rounded Data

Sai Ma[1], Jun Yan[1], and Xuebin zhang[2]

[1]University of Connecticut
[2]Environment and Climate Change Canada

November 26, 2022

Abstract

In extreme value analysis, quantization due to rounding causes biases in parameter estimation and incorrect sizes in goodness-of-fit testing. We treat rounded data as interval censored and estimate the parameters by maximizing the likelihood that accounts for interval censoring. The resulting estimator are asymptotically unbiased. Further, classic goodness-of-fit tests such as Anderson–Darling are adapted to discrete data resulted from rounding, which gives tests with correct sizes and substantial powers. Such tests have important implications in threshold selection for extreme value analyses. The performances of the estimation and goodness-of-fit are validated in a simulation study with rounded data from generalized Pareto distributions. In applications to the precipitation data of 18 eastern Washington stations, the proposed methods selected thresholds for more stations with more exceedances and, hence, more accurate return level estimations.

Extreme Value Modeling with Generalized Pareto Distributions for Rounded Data

Sai Ma[1], Jun Yan[1,2], Xuebin Zhang[3]

[1]Department of Statistics, University of Connecticut, Storrs, CT, USA.
[2]Center for Environmental Sciences and Engineering, University of Connecticut, Storrs, CT, USA.
[3]Climate Research Division, Environment and Climate Change Canada, Downsview, Ontario, Canada.

Key Points:

- The bias in the naive likelihood estimator of GPD parameters with rounded data is corrected with a true likelihood estimator.
- The incorrect size of naive goodness-of-fit tests for GPDs with rounded data is corrected.
- In batched return level estimation, the thresholds selected based on the proposed method gives lead to more accurate estimates.

Corresponding author: Jun Yan, jun.yan@uconn.edu

Abstract

In extreme value analysis, quantization due to rounding causes biases in parameter estimation and incorrect sizes in goodness-of-fit testing. We treat rounded data as interval censored and estimate the parameters by maximizing the likelihood that accounts for interval censoring. The resulting estimator are asymptotically unbiased. Further, classic goodness-of-fit tests such as Anderson–Darling are adapted to discrete data resulted from rounding, which gives tests with correct sizes and substantial powers. Such tests have important implications in threshold selection for extreme value analyses. The performances of the estimation and goodness-of-fit are validated in a simulation study with rounded data from generalized Pareto distributions. In applications to the precipitation data of 18 eastern Washington stations, the proposed methods selected thresholds for more stations with more exceedances and, hence, more accurate return level estimations.

Keywords: interval censored; likelihood estimation; quantization error; threshold selection; return level

1 Introduction

Quantization due to rounding is known to affect the results of extreme value analysis significantly in multiple ways. Precipitation data, for example, are often subject to quantization because of the limitation in the measuring devices' precision. The United States precipitation data from the Global Historical Climatology Network are available with observations recorded to the nearest hundredth of an inch. In extreme value analyses of such precipitation data with the generalized Pareto distribution (GPD), when quantization is ignored, the maximum likelihood estimators (MLE) of the GPD parameters have been reported to be biased, with more severe bias under higher quantization level relative to the data (e.g., Deidda & Puliga, 2009). Goodness-of-fit tests designed for data with no quantization, such as the Anderson–Darling test, become invalid in that the critical values of the testing statistic should be much greater under quantization (Deidda, 2007; Deidda & Puliga, 2009). That is, the tests reject when they should not. Consequently, in application to threshold selection for the peaks-over-threshold (POT) approach for extreme value analyses, all candidate thresholds could be rejected when some of them should not, and an overly high threshold could be selected when a lower one should (Langousis et al., 2016; Bader et al., 2018).

Quantized data are grouped continuous data, which have been extensively studied in the statistical literature (e.g., Heitjan, 1989; Schneeweiß et al., 2010). It has been long recognized that quantized data have a different likelihood function than that obtained when the quantization is ignored. For example, Kempthorne (1966) stated that "all observations are in fact discrete, with a grouping error specified by the scientist"; the contribution of an observed data point to the likelihood is the probability of its falling into the observed grouping intervals. Giesbrecht and Kempthorne (1976) provided an application to the three-parameter lognormal distribution. The consistency and asymptotic efficiency of the MLE follow from the multi-parameter, discrete distribution version of the general theorems of the asymptotic properties of MLEs (Kulldorff, 1957). A more advanced application is Bai et al. (2009), where a composite likelihood was used for time series models with rounded observations based on interval censoring. The method was later extended to rounded data from general dependent sequences (Zhang et al., 2010). The likelihood construction based on interval censoring for rounded data was reinvented in the engineering literature with quantized data from normal distributions (Vardeman & Lee, 2005). In extreme value analyses, Bader et al. (2018) proposed a multiple imputation approach where the rounded data are jittered multiple times with a perturbation controlled by the rounding error, and then the medians of the estimators and the testing statistics are used as estimators and testing statistics. The resulting estimator may

be a good one, but the resulting testing statistics obviously remain incompatible with the null distribution obtained for continuous data. A more elegant solution is needed.

Our contribution is a well-tested toolset for parameter estimation and goodness-of-fit test in extreme value analyses under quantization error. The parameters are estimated by maximizing the true likelihood of the observed rounded data, which is constructed based on interval censoring. When there is no quantization error, the likelihood coincides with that for continuous data. For goodness-of-fit tests, by treating the rounded data as discrete data, standard tests that are designed for continuous data are adapted, and the p-values are obtained through parametric bootstrap. In an extensive simulation study, the MLEs of the GPD parameters appeared to be unbiased; the goodness-of-fit tests held their sizes and had substantial power. When applied to threshold selection in the POT method in extreme value analyses of 18 eastern Washington stations, drastically different selection results were obtained, which led to drastic difference in point and interval estimation of return levels.

The rest of the paper is organized as follow. The problem is set up and the estimation based on the true likelihood is presented In section 2. In section 3, we adapt classic goodness-of-fit tests for GPD to the case of rounded data. A simulation study is reported in Section 4 to demonstrate the performance of the MLE and the goodness-of-fit tests. The methods are applied in Section 5 to automated threshold selection for the extreme value analyses of the precipitation of two sites in California. A discussion concludes in Section 6.

2 Likelihood Estimation

Without loss of generality, consider a GPD with location zero and distribution function

$$F(x;\sigma,\xi) = \begin{cases} 1-\left[1+\frac{\xi x)}{\sigma}\right]^{-1/\xi}, & \xi \neq 0; \\ 1-\exp\left(-\frac{x}{\sigma}\right), & \xi = 0, \end{cases}$$

where $\sigma \in [0,\infty)$ and $\xi \in (-\infty,\infty)$ are scale and shape parameters, respectively. The support of the distribution is $x \geq \mu$ if $\xi \geq 0$ and $0 \leq x \leq -\sigma/\xi$ if $\xi < 0$. Let $\mathbf{X} = (X_1,...,X_n)$ be random samples of size n from $F(\cdot;\sigma,\xi)$. If there were no rounding, we would observe $\mathbf{x} = (x_1,...,x_n)$. With rounding, however, we only observe a rounded version of $\mathbf{x}$ subject to a known rounding level $\delta > 0$. Each observed data point is rounded to the nearest multiple of δ. Let $\mathbf{x}^* = (x_1^*,...,x_n^*)$ be the observed rounded sample. We have $x_i^* = \delta\lceil x_i/\delta - 0.5\rceil$ for $i = 1,2,...,n$, where $\lceil t\rceil$ is the ceiling integer of t, that is, the least integer greater than or equal to t. Our task is to estimate σ and ξ based on the observed data $\mathbf{x}^*$.

The contribution of observation x_i^* to the likelihood is

$$\Pr(X_i \in [(x_i^*-\delta/2)^+, x_i^*+\delta/2)) = F\left(x_i^*+\frac{\delta}{2}\right) - F\left(\left(x_i^*-\frac{\delta}{2}\right)^+\right), \tag{1}$$

where $t^+ = \max(0,t)$. Note that $x_i^*+\delta/2$ can be beyond the support of the distribution if $\xi < 0$, but it has no influence to the MLE since $F(x_i^*+\frac{\delta}{2}) = 1$ in this case. The loglikelihood function is thus

$$\ell^*(\sigma,\xi;\mathbf{x}^*) = \sum_{i=1}^{n} \log\left\{F\left(x_i^*+\frac{\delta}{2}\right) - F\left(\left(x_i^*-\frac{\delta}{2}\right)^+\right)\right\}. \tag{2}$$

The maximizer of the loglikelihood (2), $(\hat{\sigma},\hat{\xi})$ is the MLE of (σ,ξ). As long as $\xi > -0.5$, the MLE is asymptotically unbiased and normally distributed, with variance being the inverse of the Fisher information matrix (Smith, 1985).

The likelihood contribution in (1) is the key to construct the correct likelihood based on the rounded data. A naive treatment is to ignore the rounding, pretending that the rounded data are continuous observations. In that case, the likelihood contribution of x_i^* is $f(x;\sigma,\xi)$, where f is the probability density function of $F(x;\sigma,\xi)$. As $\delta \to 0$, the contribution in (1) divided by δ converges to $2f(x_i^*;\sigma,\xi)$ for $x_i^* > 0$ and to $f(0;\sigma,\xi)$ for $x_i^* = 0$. Therefore, the naive MLE is similar to the true MLE only δ is small relative to the scale of the distribution. As will be shown, when δ increases, the bias of the naive MLE increases while the true MLE remains unbiased. Further, for large samples, the variance of the true MLE can be reasonably well estimated by the inverse of the Fisher information matrix.

3 Goodness-of-fit Test

Let $F_\delta(\cdot;\sigma,\xi)$ be the discretized version of $F(\cdot;\sigma,\xi)$ rounding level $\delta > 0$. For a given parameter vector (σ,ξ), $F_\delta(t;\sigma,\xi)$ is the distribution function of a discrete random variable with support $t \in \{0,\delta,2\delta,\cdots\}$; at each point in the support, we have $F_\delta(t;\sigma,\xi) = F(t+\delta/2;\sigma,\xi)$. The goodness-of-fit test is to test

$$H_0 : \mathbf{x}^* \text{ is a random sample from distribution function } F_\delta(\cdot|\sigma,\xi) \text{ for some } (\sigma,\xi). \tag{3}$$

3.1 Chi-Squared Test

The first test to be considered it the chi-squared (CS) test (Pearson, 1900), which applies to both continuous and discrete data. Suppose that the data are grouped into k bins. The testing statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the observed count and E_i is the expected count in the ith bin, $i = 1,\ldots,k$. In practice, each E_i is usually set to be no smaller than 5. It is critical to note that the expected count E_i is calculated based on the MLE $(\hat{\sigma},\hat{\xi})$ of the parameters of the hypothesized GPD. This MLE should be the true MLE based on interval censoring instead of the naive MLE.

To assess the significance of the test, the null distribution of the testing statistic is needed, which is not as simple as many think. Only when the parameters are estimated by minimizing the test statistic does the test statistic under H_0 follows a χ^2_{k-p-1} distribution (Fisher, 1922), where $p = 2$ is the number of parameters in the GPD case. The MLE in general is not the same as that estimate. Therefore, the null distribution of the testing static is between χ^2_{k-p-1} and χ^2_{k-1} (Chernoff & Lehmann, 1954). We approximate the p-value of the observed statistic through parametric bootstrap, where each bootstrap sample is generated from the fitted distribution (Tibshirani & Efron, 1993, Chapter 21).

In our numerical studies, the bins were chosen so that the fitted relative frequency for each bin is at least 0.05; otherwise, it is merged with a neighboring bin.

3.2 Tests Adapted from the Continuous Case

The Kolmogorov–Smirnov (KS), Cramér–von Mieses (CvM) and Anderson–Darling (AD) tests are more powerful than the CS test for continuous data. The test statistics of these tests are obtained by comparing the empirical distribution function with a fitted parametric distribution function. The fitted parameter values should be the true MLE instead of the naive MLE. With a continuous null distribution, there is no tie in the sample, and the distribution of these test statistic does not depend on the hypothesized distribution. For a discrete hypothesized distribution, the calculation of the test statistic

remains the same, but the null distribution of the test statistic does depend on the particular hypothesized distribution (Conover, 1972; Choulakian et al., 1994). When the hypothesized discrete distribution contains no unknown parameters, these tests have been adapted (Arnold & Emerson, 2011) and implemented in popular software packages (Arnold & Emerson, 2013). In our case, since the parameters are estimated, the p-value returned from the software cannot be used. Again, we use parametric bootstrap to approximate the p-value of each of the KS, CvM, and AD test.

One caveat in applying the adaptation is that the tests of Arnold and Emerson (2011) are for discrete null distribution with finite support. For GPD, the support is not finite if $\xi > 0$. Fortunately, the test statistic only depends on the range of the observed data. We simply truncate the support by the largest observed value $x^*_{(n)}$.

As to be shown in the simulation study, these tests have higher power than the CS test in this application we conjectured.

4 Numerical Study

Base on the settings of Deidda and Puliga (2009), a numerical study was carried out to validate the performance of the proposed estimator and the goodness-of-fit tests. The true parameters (σ, ξ) were set to be $\sigma \in \{7, 12\}$ and $\xi \in \{-0.2, 0, 0.3\}$. The rounding level δ was set to be on a grid from 0 to 4 with increment 0.5, where $\delta = 0$ meant no rounding. In total, we had $2 \times 3 \times 9 = 54$ settings. For each setting, we generated 1000 datasets with sample size $n = 500$. For comparison, the naive MLE obtained with the quantization ignored was included using the implementation in the `R` package `eva` (Bader & Yan, 2018). The true MLE was obtained by the `optim()` function in `R` with the default Nelder–Mead algorithm, using the naive MLE as the initial value. The estimation for all the datasets converged. In the sequel, the true MLE based on interval censoring is denoted as MLE-IC, while the naive MLE is denoted as MLE-N. The MLE-N is the same as the true MLE only when $\delta = 0$.

4.1 Estimation

Figure 1 shows bias of the MLE-N and MLE-IC for each combination of (σ, ξ) as a function of δ. The MLE-N is positively biased for σ and negatively biased for ξ for $\delta > 0$. For example, when $\delta = 4$, the bias of $\hat{\sigma}$ in the setting of $(\sigma, \xi) = (7, 0.3)$ is almost 5, which is a quite large bias. The magnitude of the bias of MLE-N is clearly increasing as δ increases. For a given scale parameter σ, the MLE-N bias is larger for higher in both parameters for higher ξ. For a given shape parameter ξ, the relative bias of MLE-N in both parameters is higher for lower σ. These results for the MLE-N are similar to earlier studies (Deidda & Puliga, 2009). In contrast, the bias of the MLE-IC is virtually zero regardless of the values of rounding level δ or parameters (σ, ξ), as expected from the asymptotic properties of MLE.

The square root of the mean squared errors (RMSE) are compared in Figure 2. The pattern of the RMSE of the MLE-N in response to δ, σ, and ξ is similar to pattern of the magnitude of the bias in Figure 1. This is because the mean squared error of the MLE-N is dominated by its bias. The RMSE of the MLE-IC, on the contrary, is dominated by its variance, which increases only slightly as δ increases. This is expected as higher rounding level means less information, but the rate of the increase almost flat compared to that in the case of MLE-N. We also investigated the estimated standard errors of the MLE-IC, whose averages were in close agreement with the empirical standard errors (not shown). This suggests that the uncertainty in MLE-IC, which is necessary in making statistical inferences, can be well estimated by inverting the Fisher information matrix for the sample size $n = 500$ in this study.

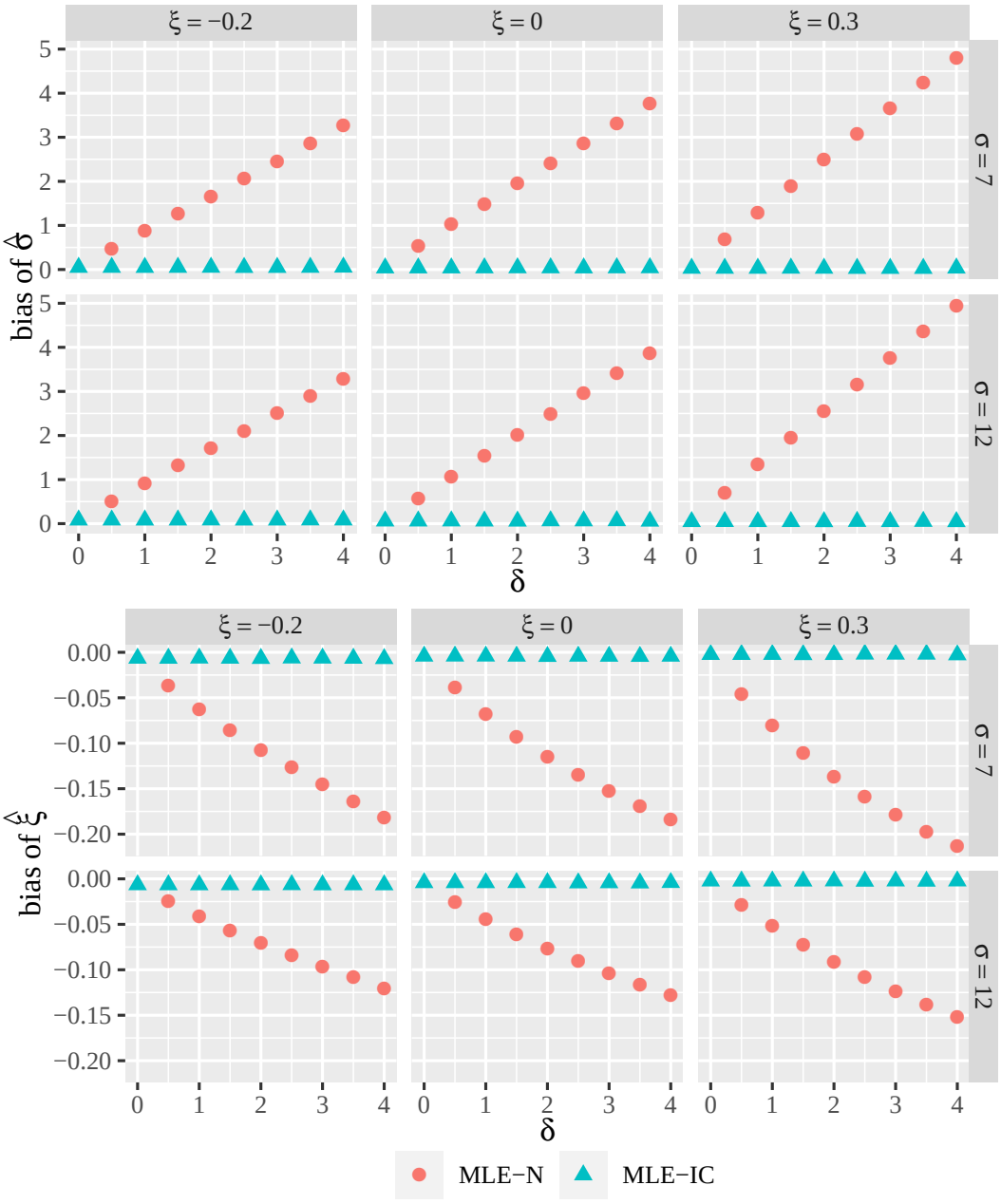

Figure 1. Bias of naive MLE and MLE-IC. Above: bias plots of shape estimators $\hat{\sigma}$. Below: bias of scale estimators $\hat{\xi}$.

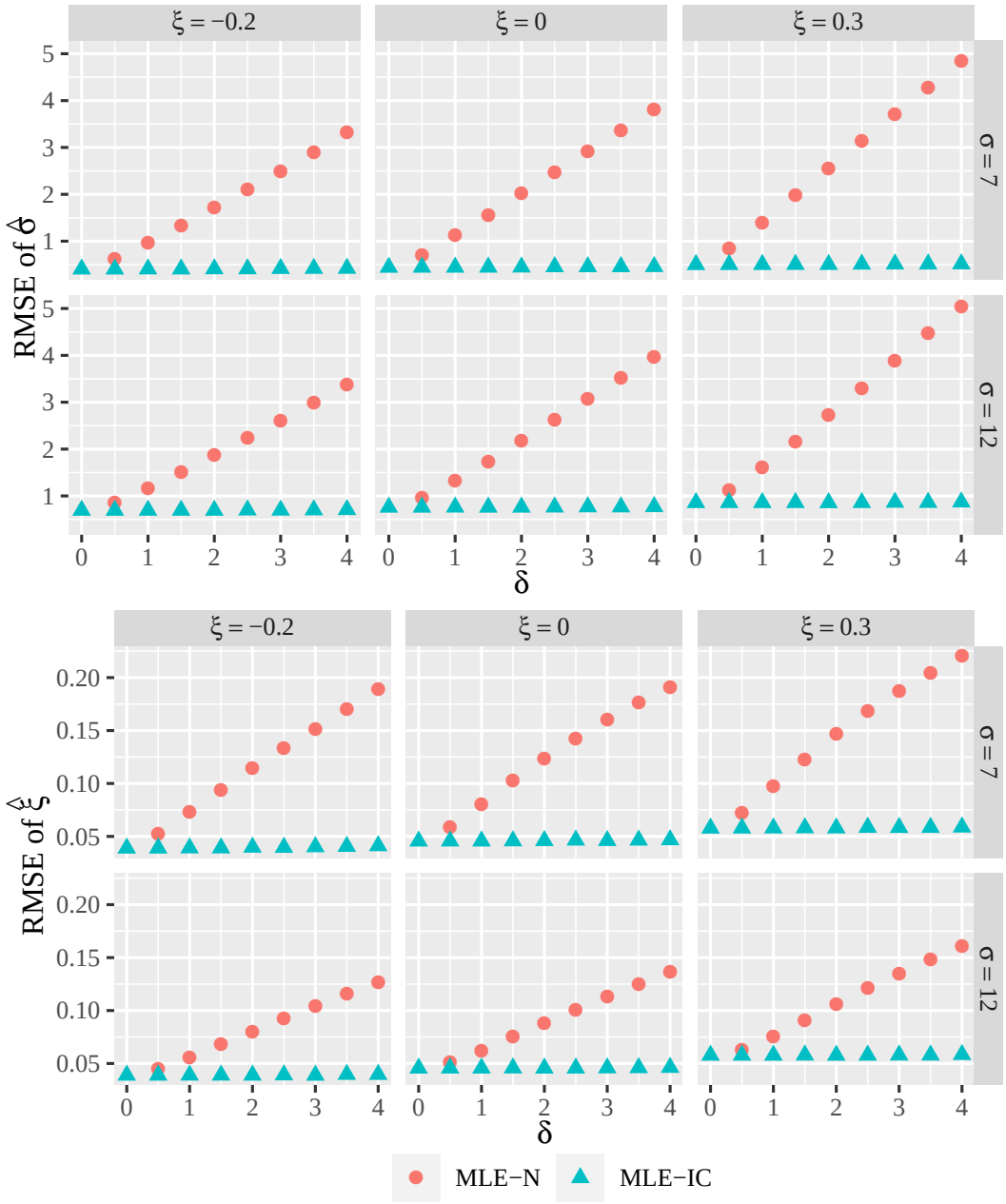

Figure 2. RMSE plots of naive MLE and MLE-IC. Above: RMSE plots of shape estimators $\hat{\sigma}$. Below: RMSE plots of scale estimators $\hat{\xi}$.

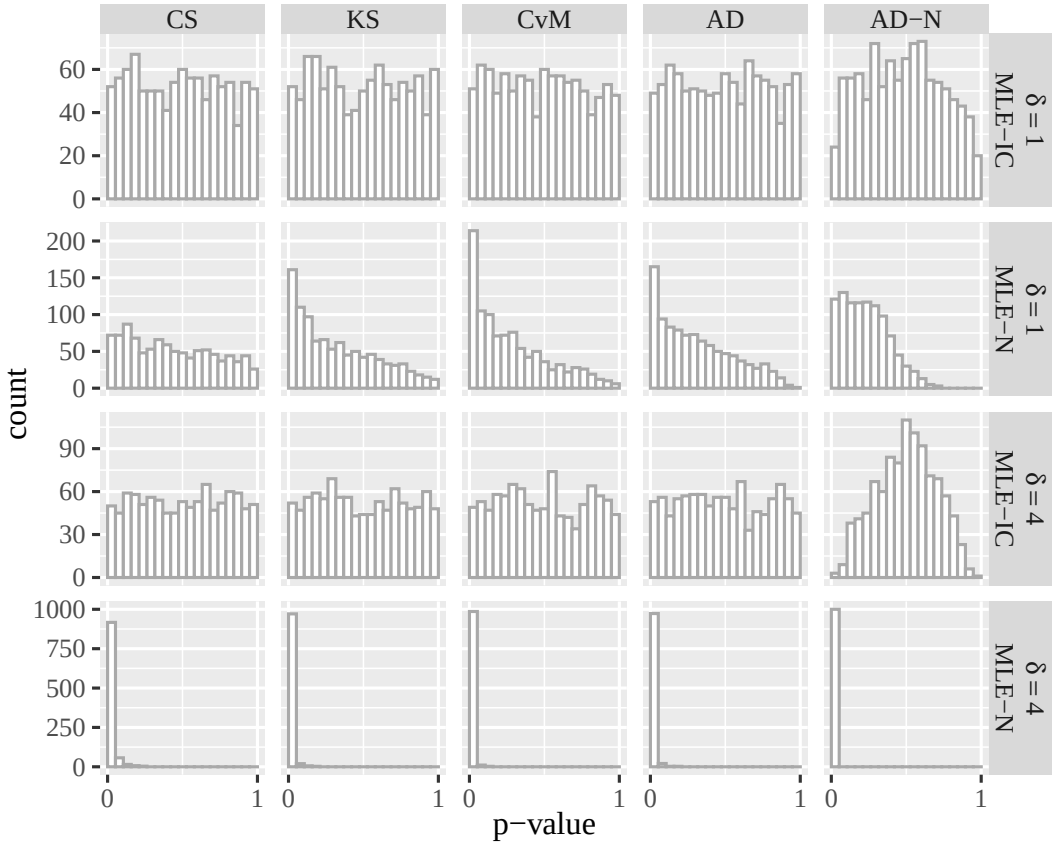

Figure 3. Some goodness-of-fit test for rounded-off samples. Samples were generated by GPD with $(\sigma, \xi) = (7, 0.3)$, and rounded off to $\delta = 1$ (above two rows) and $\delta = 4$ (below two rows). Five goodness-of-fit tests and two estimation methods were used here.

We also compared the estimates of the 25-, 50-, 100-, and 200-year return levels using the two methods. There is virtually no bias from the MLE-IC estimates regardless of the level of δ. The MLE-N estimates have larger bias for larger δ, and the return levels are over estimated for shorter return periods but underestimated for longer return periods. The MSEs of the two methods shows that the MLE-IC is much preferred to the MLE-N, especially when the bias of the latter is huge. The figures of the biases and MSEs of the return level estimates are presented in the Supplementary Materials.

4.2 Goodness-of-Fit Test

4.2.1 Size

Goodness-of-fit tests were performed on each of the 54 settings in the last section. We compared 5 tests. The first four are those presented in Section 3: CS, KS, CvM, and AD. The last one is a naive AD test, applied with the quantization ignored. That is, the rounded observed data were treated as if they were continuous data. This test is denoted as AD-N. For all five tests, we also performed a version where the fitted distribution was based on MLE-N instead of the true MLE-IC. This version helps to show how seriously wrong the tests based on MLE-N can be.

Figure 3 shows that histogram of the 1000 p-values from tests for the settings with $(\sigma, \xi) = (7, 0.3)$ and $\delta \in \{1, 4\}$. The results from other settings convey the same mes-

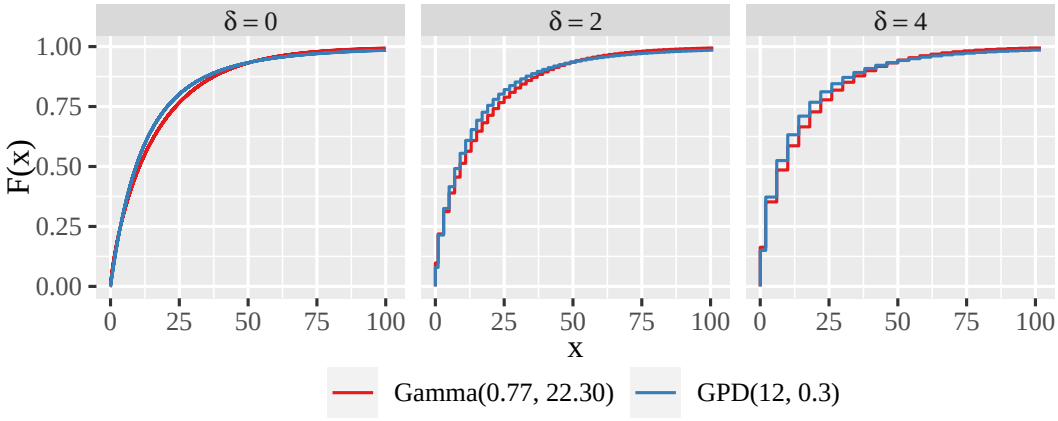

Figure 4. The CDF plots. Left: continuous CDF. Middle: rounding is 2. Right: rounding is 4.

sages and, hence, are now shown. Since the data were indeed generated from GPD distributions before being rounded, we expect that the p-values from the 1000 replicates to be uniformly distributed over $(0, 1)$. The expected histograms are observed for four proposed tests, CS, KS, CvM, and AD, when the fitted distribution were evaluated with MLE-IC; and this is true regardless of the rounding level δ. Using MLE-N in place of MLE-IC causes these tests to have more smaller p-values and to reject more often than desired. The over-rejection gets even worse as δ increases from 1 to 4. The AD-N test over-rejects when the fitted distribution was evaluated at MLE-N. When MLE-IC was used, it performs approximately as desired only when $\delta = 0.5$ (not shown); the deviation of the histogram from the standard uniform distribution gets more severe for larger δ.

Because of the parametric bootstrap process in each test, this simulation study was very time-consuming. The computing time of the CS test with a fixed number of partitions does not increase as δ decreases. For the KS, CvM, and AD tests, however, the the most time-consuming setting were $\delta = 0.5$, because the observed data has much more unique points. The whole study was run on the high performance cluster of the University of Connecticut.

4.2.2 Power

For the four tests that performed as desired under H_0, we compared their powers when the data were not generated under H_0. Specifically, we considered a gamma distribution which is closest to a GPD distribution with $(\sigma, \xi) = (12, 0.3)$ in terms of the Kullback–Leibler distance. The parameters of this gamma distribution were identified through a Monte Carlo approximation as $(\alpha, \beta) = (0.77, 22.30)$, where α and β are the shape and scale parameters, respectively. The rounding level was set to be $\delta \in \{2.0, 2.5, 3.0, 3.5, 4.0\}$. Figure 4 shows that the distribution functions of Gamma$(0.77, 22.30)$ and GPD $(12, 0.3)$ under no discretization ($\delta = 0$) and under rounding level $\delta = 2$ and $\delta = 4$. By construction, the two distributions are similar, making the goodness-of-fit test for a GPD distribution with data from the gamma distribution a challenging problem. Two sample sizes were considered, $n \in \{250, 500\}$. For each setting, 1000 datasets were generated. For each dataset, the four goodness-of-fit tests were applied, and H_0 was rejected with significance level 0.05.

Figure 5 presents the empirical powers of the fours tests. The AD test has the highest power, followed by the CvM test and then the KS test. The CS test has the lowest power. The AD test's power is about 60% when $\delta = 2$, twice as high as the CS test,

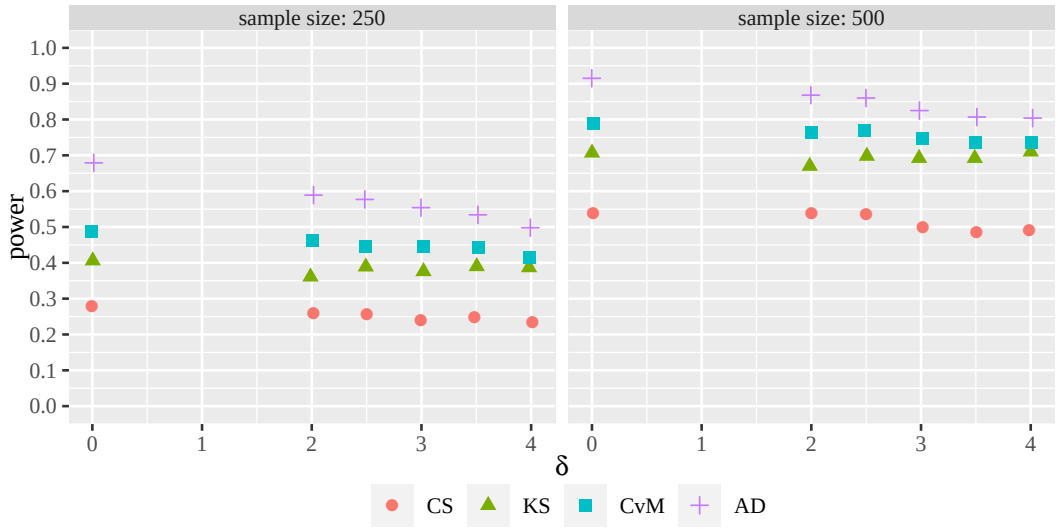

Figure 5. Power of five goodness-of-fit tests for GPD $(12, 0.3)$ with nominal size 0.05. Alternative distribution is Gamma $(0.77, 22.30)$. Left: sample size is 250. Right: sample size is 500.

for sample size 250. At sample size 500, the AD test has power about 87%, while the CS test's power is only about 55%. Given how similar the two distributions are, the power of the AD test is substantial. There is a clear decreasing trend of the power as δ increases for all the test except the KS test, which is expected since higher δ means more information loss. As the sample size increases from 250 to 500, all tests have higher power.

5 Application to Annual Maximum of Daily Precipitations

The proposed method was applied to model daily precipitation data of 18 monitoring stations from 1969 to 2018 in the eastern part of the Washington State. This area is known to be dry with much less rainfall than the western part due to the "rain shadow" effect of the Cascade Mountains. From the simulation studies, we expect more obvious differences between inferences based on MLE-N and inferences based on MLE-IC in areas with less precipitations. Daily precipitation data for the 18 stations were obtained from the Global Historical Climatology Network (Menne, Durre, Vose, et al., 2012). The same data preparation as Bader et al. (2018) was employed. Since most of the precipitations occur in winter, we only used the precipitation data in winter (from November to next March). The total number of winter days from 1969 and 2018 was 7512. For each site, we chose 15 candidate thresholds to test by taking the 70th to 98the percentiles in increments of 2 percent. The data were recorded to the nearest hundredth of an inch. With inch being the unit, we have $\delta = 0.01$.

The threshold selection for the POT approach in Bader et al. (2018) relies critically on the goodness of fit test of the GPD distribution at each candidate threshold. We focus on the AD test as it has the highest power in the simulation study. The result of threshold selection from the sequential AD tests has three possibilities: 1) no threshold is selected from either MLE-IC or MLE-N; 2) a threshold is selected from MLE-IC but not from MLE-N; 3) a threshold is selected from both methods but the one from MLE-IC is lower than the one from MLE-N. The second and third possibilities show the advantage of the MLE-IC in selecting threshold with more exceedances and, hence, higher efficiency in inferences.

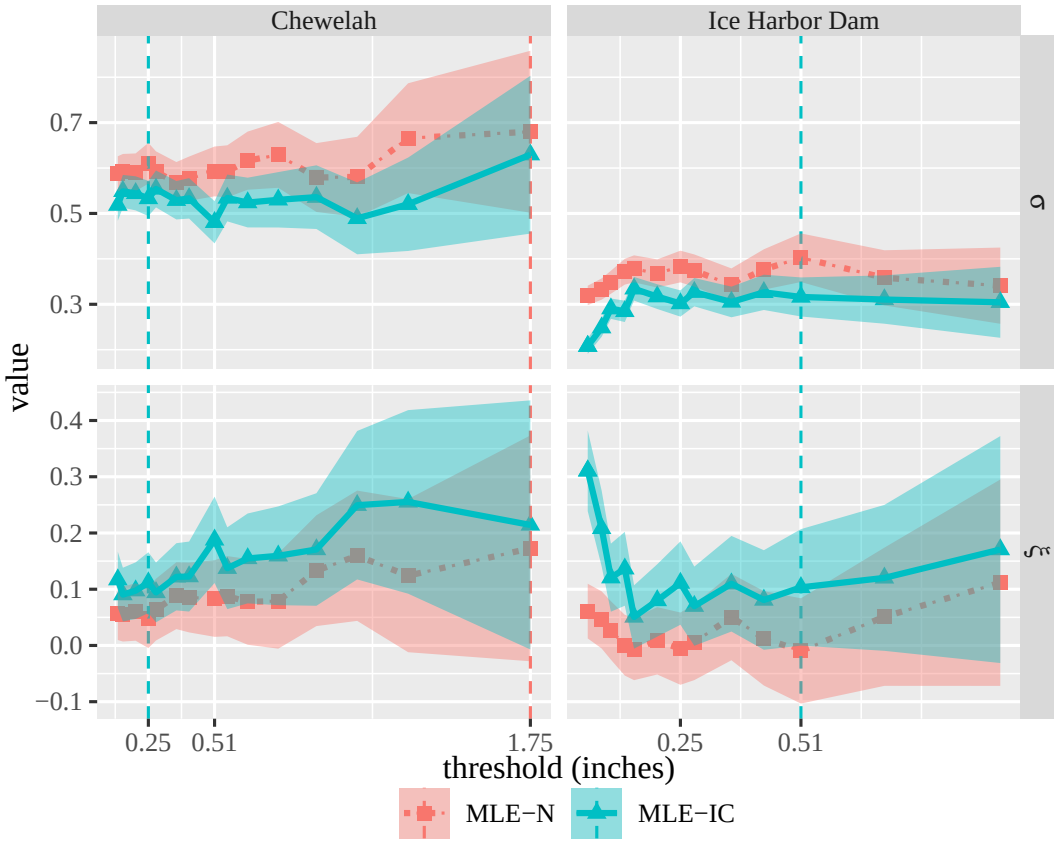

Figure 6. GPD parameter estimates and 95% confidence intervals. The MLE-N and MLE-IC are obviously different in both point estimates and uncertainty.

To illustrate the differences from the two tests, consider two stations, one at Chewelah and the other at Ice Harbor Dam. The average yearly winter precipitation was 29.13 inches in Chewelah and 14.99 inches in Ice Harbor Dam. The total number of winter precipitation days was 2533 and 2778, respectively. A table in the Supplementary Materials summarizes the candidate thresholds and the corresponding number of exceedances at the two sites. Figure in the supplementary materials shows the p-values at the 15 candidate thresholds at the two sites before and after the ForwardStop adjustment using the AD test as in Bader et al. (2018). At Chewelah, the tests based on MLE-N selects threshold 1.75 with 135 exceedances; the tests based on MLE-IC selects 0.25 as the threshold with 1596 exceedances. The number of exceedances from MLE-IC is almost 12 times of that from MLE-N. At Ice Harbor Dam, the tests based on MLE-N rejected all candidate thresholds while the tests based on MLE-IC selects threshold 0.51 with 430 exceedances. The threshold selected by the tests based on MLE-IC makes return level estimation possible.

The statistical inference results are largely affected by the threshold. Figure 6 shows the parameter estimates $\hat{\sigma}$ and $\hat{\xi}$ and their 95% confidence intervals for the two sites if a threshold was selected. At Chewelah, with a lower threshold and more exceedances, the parameters are estimated with a much narrower confidence intervals by MLE-IC than those by MLE-N. At Ice Harbor Dam, the tests based on MLE-IC shows that the tail of the annual maximum daily can be modeled by a GPD, allowing estimation of return levels which would otherwise be impossible if MLE-N were used. A figure in the Supplementary Materials presents the estimated 25-, 50-, 100-, and 200-year return levels

along with 95% confidence intervals constructed from profile likelihood for the two sites. The return level estimates from the two methods at Chewelah are very different, with the confidence interval based on MLE-N about 10 times wider than that based on MLE-IC. The return level estimates at Ice Harbor Dam are smaller than those at Chewelah as expected, with similarly narrow confidence intervals. The plots for the return level estimates are in the Supplementary Materials.

We performed a batch analysis of the return levels on the 18 sites; the selected thresholds and the corresponding number of exceedances are summarized in the Supplementary Materials. Among all 18 stations in eastern Washington, 9 stations had thresholds selected via MLE-IC, but only 2 station had threshold selected via MLE-N. The stations with a threshold selected via MLE-N are a subset of those via MLE-IC. At stations where both methods yielded a threshold, the numbers of exceedances from MLE-IC are much greater than those from MLE-N, which has important implications on inferences on the GPD parameters and the return levels. Although Bader et al. (2018) reported that the jittering method can also fix some issues, it is an ad-hoc, partial fix while the MLE-IC provides a complete and desired solution. If the same analysis were done on all the sites analyzed in Bader et al. (2018), we would expect that some dry sites with no threshold selected may have a threshold selected, and some sites with a high threshold selected may have a lower threshold selected. Consequently, return levels at many sites may be estimated much more accurately.

6 Discussion

Bias in parameter estimation and over-rejection in goodness-of-fit tests have been documented as consequences of quantization error in extreme value analyses (Deidda, 2007; Deidda & Puliga, 2009) but without satisfying solutions. Our MLE based on interval censoring and goodness-of-fit tests adapted from continuous distributions to discrete distributions provide a solid, feasible approach to the problem. The inferences based on the asymptotic normality of the MLE appear to be valid for the sample size investigated. The method has wide applications in extreme value analyses of precipitation or temperature, which could lead to quite different results than those obtained otherwise. When the generalized extreme distribution is used or when some parameters incorporate covariates, the interval censoring framework can be applied too.

The correction of the method to the naive method depends on the rounding level relative to the scale parameter. For example, if mm instead of inch is the unit of precipitations in the illustratie example, the rounding level would be by 25.4. Nonetheless, the change of unit would only changed the estimated scale parameter of GPD by the same multiplier; the estimated shape parameter and p-value of the goodness-of-fit should remain the same. The method requires that the rounding level δ is known. It is applicable to datasets with multiple rounding levels; for example, later data may be more precise than earlier data. Although the AD test for discrete data has the hightest power, it is much more computing intensive than the CS test, especially when δ is small relative to σ and $\xi > 0$, in which case, the support of the discretized distribution has a large number of points. A faster alternative would be of interest.

The impact of correcting the bias with MLE-IC is greatest at locations with less precipitations. Our illustration focused on the 18 eastern Washington stations which are known to be much drier than those to the west of the mountains. In batch studies such as Bader et al. (2018), we expect more drastic differences if MLE-IC were used in place of MLE-N, in terms of the number of stations with a threshold selected, the number of exceedances, and the resulting point and interval estimate of various return levels.

Acknowledgments

The data used in the Application Section are available from the Global Historical Climatology Network in Menne, Durre, Korzeniewski, et al. (2012).

J. Yan's research was partially supported by the NSF grant DMS 1521730.

References

Arnold, T. B., & Emerson, J. W. (2011). Nonparametric goodness-of-fit tests for discrete null distributions. *R Journal*, *3*(2), 34–39.

Arnold, T. B., & Emerson, J. W. (2013). dgof: Discrete goodness-of-fit tests [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=dgof (R package version 1.2)

Bader, B., & Yan, J. (2018). eva: Extreme value analysis with goodness-of-fit testing [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=eva (R package version 0.2.5)

Bader, B., Yan, J., & Zhang, X. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *The Annals of Applied Statistics*, *12*(1), 310–329.

Bai, Z., Zheng, S., Zhang, B., & Hu, G. (2009). Statistical analysis for rounded data. *Journal of Statistical Planning and Inference*, *139*(8), 2526–2542.

Chernoff, H., & Lehmann, E. (1954). The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *The Annals of Mathematical Statistics*, *25*(3), 579–586.

Choulakian, V., Lockhart, R. A., & Stephens, M. A. (1994). Cramér-von Mises statistics for discrete distributions. *Canadian Journal of Statistics*, *22*(1), 125–137.

Conover, W. J. (1972). A Kolmogorov goodness-of-fit test for discontinuous distributions. *Journal of the American Statistical Association*, *67*(339), 591–596.

Deidda, R. (2007). An efficient rounding-off rule estimator: Application to daily rainfall time series. *Water resources research*, *43*(12), W12405.

Deidda, R., & 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*(10–12), 626–634.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, *85*(1), 87–94.

Giesbrecht, F., & Kempthorne, O. (1976). Maximum likelihood estimation in the three-parameter lognormal distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, *38*(3), 257–264.

Heitjan, D. F. (1989). Inference from grouped continuous data: A review. *Statistical Science*, *4*, 164–179.

Kempthorne, O. (1966). Some aspects of experimental inference. *Journal of the American Statistical Association*, *61*(313), 11–34.

Kulldorff, G. (1957). On the conditions for consistency and asymptotic efficiency of maximum likelihood estimates. *Scandinavian Actuarial Journal*, *1957*(3-4), 129–144.

Langousis, A., Mamalakis, A., Puliga, M., & 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*(4), 2659–2681.

Menne, M. J., Durre, I., Korzeniewski, B., McNeal, S., Thomas, K., Yin, X., ... Houston, T. G. (2012). *Global historical climatology network—daily (GHCN-daily), Version 3.* (NOAA National Climatic Data Center) doi: 10.7289/V5D21VHZ

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

overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, *29*(7), 897–910.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *50*(302), 157–175.

Schneeweiß, H., Komlos, J., & Ahmad, A. S. (2010). Symmetric and asymmetric rounding: A review and some new results. *AStA Advances in Statistical Analysis*, *94*(3), 247–271.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, *72*(1), 67–90.

Tibshirani, R. J., & Efron, B. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Vardeman, S. B., & Lee, C.-S. (2005). Likelihood-based statistical estimation from quantized data. *IEEE Transactions on Instrumentation and Measurement*, *54*(1), 409–414.

Zhang, B., Liu, T., & Bai, Z. (2010). Analysis of rounded data from dependent sequences. *Annals of the Institute of Statistical Mathematics*, *62*(6), 1143–1173.

Supplementary Materials to Extreme Value Modeling with Generalized Pareto Distributions for Rounded Data

Sai Ma[1], Jun Yan[1,2], Xuebin Zhang[3]

[1]Department of Statistics, University of Connecticut, Storrs, CT, USA.
[2]Center for Environmental Sciences and Engineering, University of Connecticut, Storrs, CT, USA.
[3]Climate Research Division, Environment and Climate Change Canada, Downsview, Ontario, Canada.

1 Return Level Plots in the Simulation

Figure 1 shows the biases of the estimates of the 25-, 50-, 100-, and 200-year return levels from the MLE-IC and MLE-N. The corresponding MSEs are shown in Figure 2.

2 Supplementary Materials to Application

Table 1 summarizes the candidate thresholds and the corresponding number of exceedances at the two stations, one at Chewelah and the other at Ice Harbor Dam.

Figure 3 shows the p-values at the 15 candidate thresholds at the two sites before and after the ForwardStop adjustment using the AD test as in Bader et al. (2018).

Figure 4 shows the 25-, 50-, 100-, 200-year return levels and confidence intervals at the thresholds. The return levels are totally different between the two methods. Since

Corresponding author: Jun Yan, jun.yan@uconn.edu

Table 1. The candidate thresholds and number of exceedances of Chewelah station and Ice Harbor Dam station.

	Chewelah		Ice Harbor Dam	
percentiles (%)	candidate thresholds (inches)	number of exceedances	candidate thresholds (inches)	number of exceedances
70	0.13	1984	0.03	2141
72	0.15	1868	0.05	1831
74	0.20	1722	0.05	1831
76	0.25	1596	0.08	1626
78	0.28	1491	0.10	1496
80	0.36	1331	0.13	1320
82	0.41	1218	0.15	1236
84	0.51	1067	0.20	1089
86	0.56	939	0.25	920
88	0.64	828	0.28	858
90	0.76	677	0.36	718
92	0.91	526	0.43	559
94	1.07	412	0.51	430
96	1.27	289	0.69	281
98	1.75	135	0.94	141

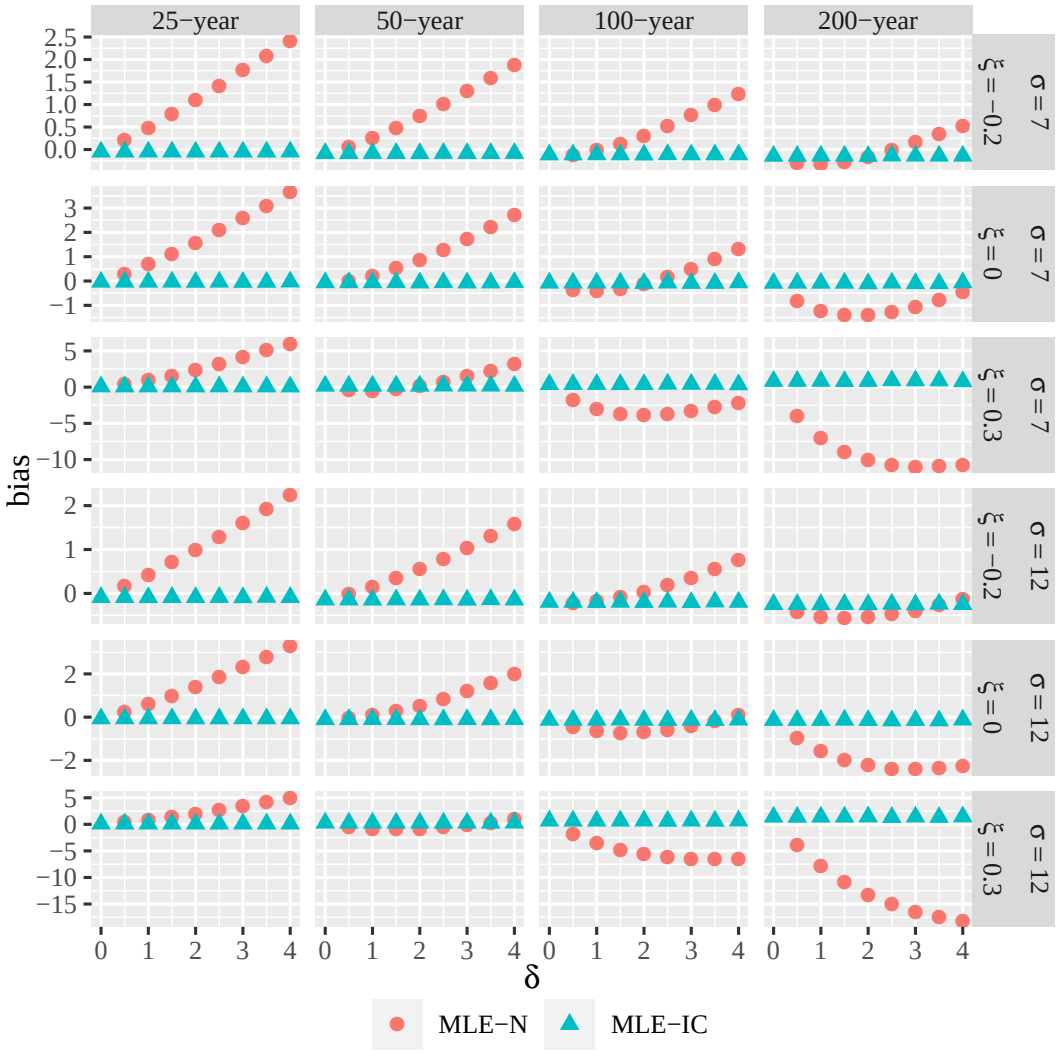

Figure 1. Bias of 25-, 50-, 100-, 200-year return level. The continuous scenarios were obtained by regular MLE. The rounded-off scenarios were obtained by naive MLE and MLE-IC.

MLE-IC has smaller estimated shape parameter and larger estimated scale parameter than MLE-N, it always has larger return level than MLE-N. Also, MLE-IC has smaller confidence interval than MLE-N.

Table 2 summarizes the selected thresholds and the corresponding numbers of exceedances.

References

Bader, B., Yan, J., & Zhang, X. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *The Annals of Applied Statistics*, *12*(1), 310–329.

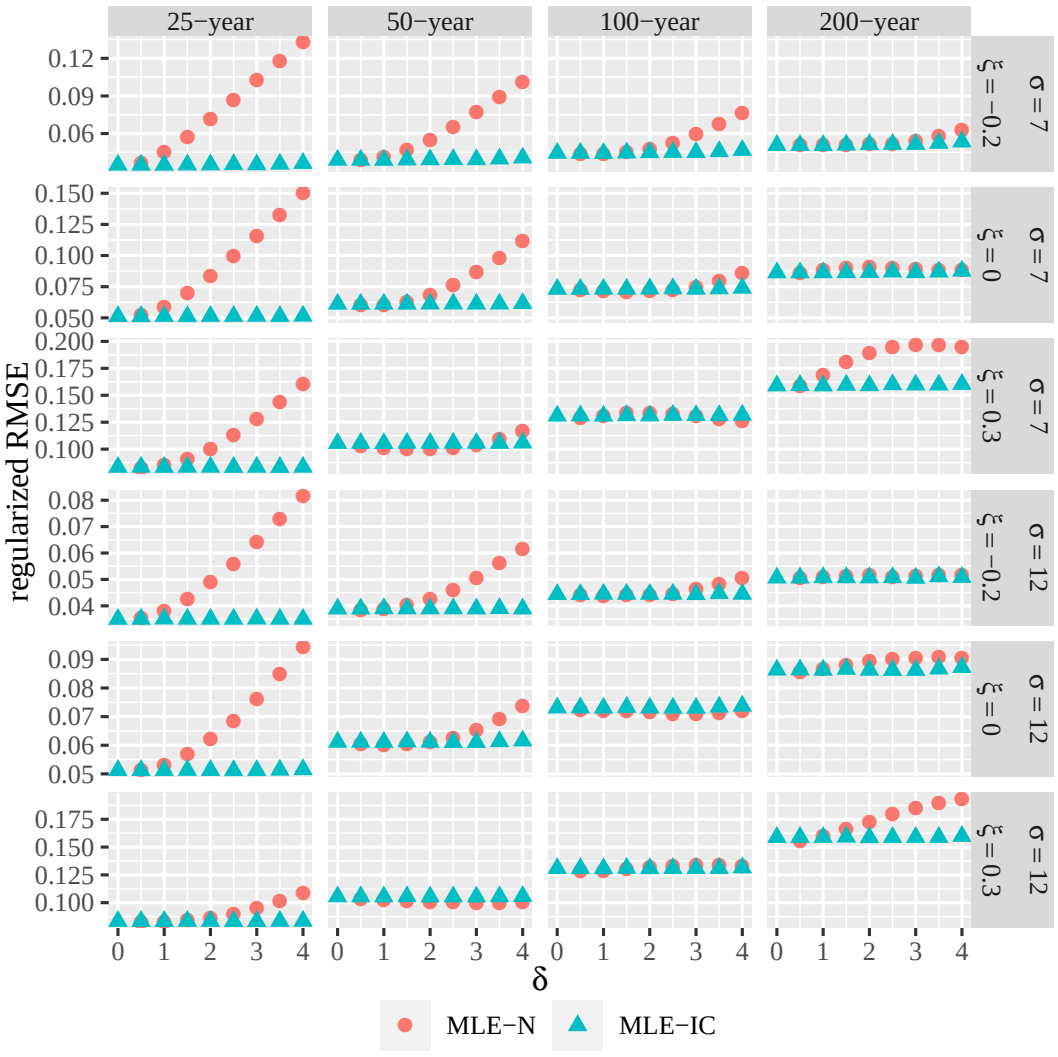

Figure 2. Regularized RMSE of 25-, 50-, 100-, 200-year return level. The continuous scenarios were obtained by regular MLE. The rounded-off scenarios were obtained by MLE-N and MLE-IC.

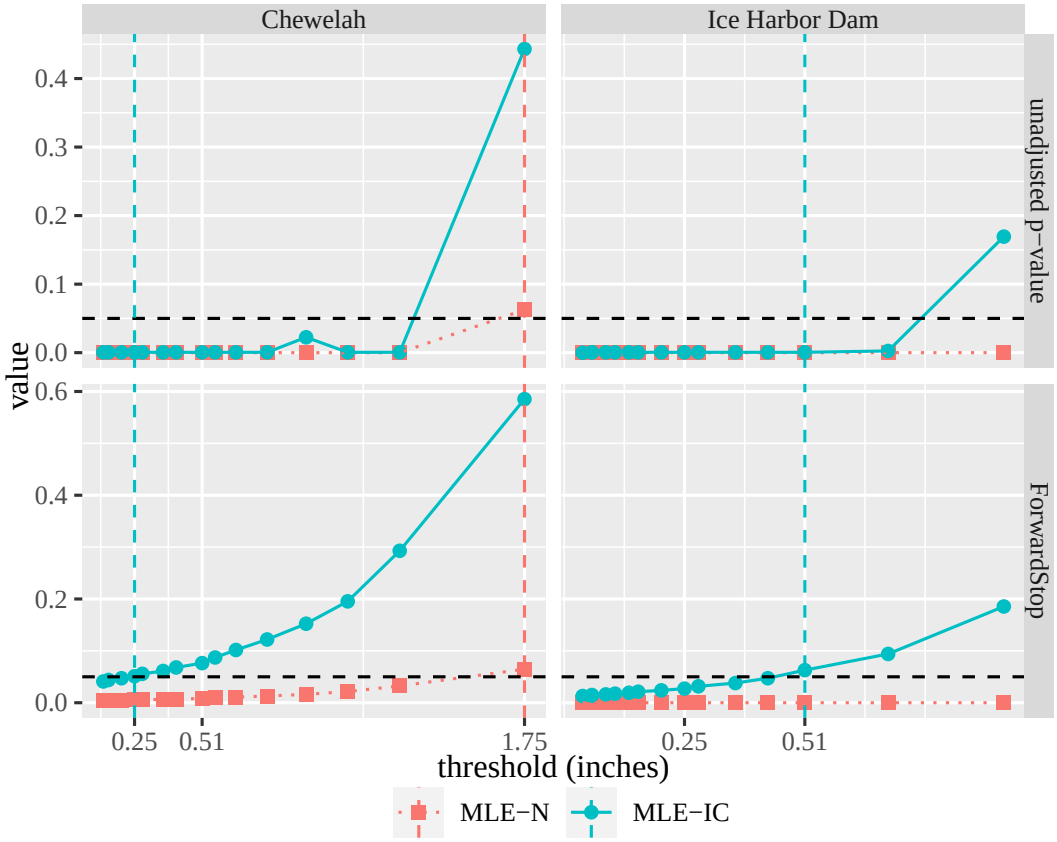

Figure 3. P-values before and after the ForwardStop adjustment. Those based on MLE-IC are higher than those based on MLE-N, so the selected thresholds from MLE-IC are smaller.

Table 2. The summary of 18 monitoring stations from 1969 to 2018 in the eastern part of the Washington State.

Station	MLE-N		MLE-IC	
	threshold	number of exceedances	threshold	number of exceedances
Chewelah	1.75	131	0.25	1596
Coulee Dam 1 SW				
Davenport				
Harrington				
Ice Harbor Dam			0.51	430
Lacrosse				
Mill Creek Dam			0.08	2190
Newport			1.30	411
Odessa				
Pomeroy			1.40	141
Pullman 2 NW			0.79	625
Republic				
Ritzville 1 SSE			0.79	310
Rosalia				
St. John				
Whitman Mission	1.12	143	0.66	425
Wilbur				
Spokane Intl AP			0.28	1515

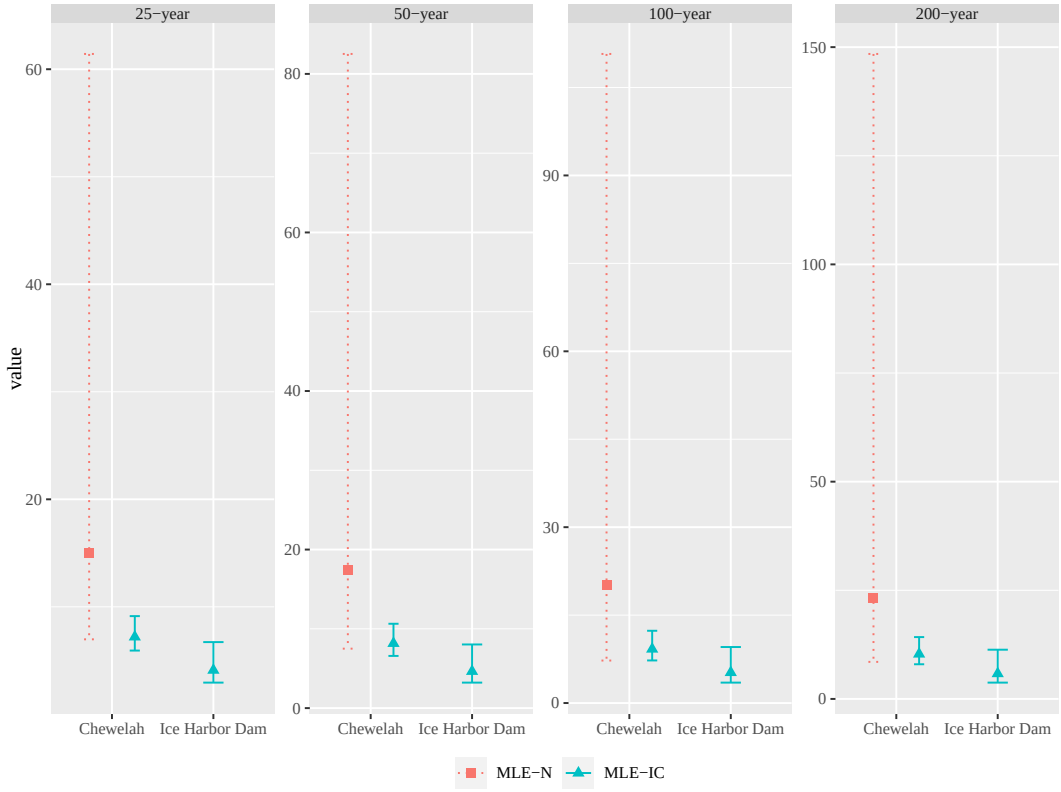

Figure 4. 25-, 50-, 100-, 200-year return level and confidence interval at the threshold of each station.