The New England Journal of Statistics in Data Science Volume 0, 1–12 (2026)
DOI: https://doi.org/10.51387/26-NEJSDS101

Extreme Value Modeling with Generalized Pareto Distributions for Rounded Data

Sai MA, Jun YAN*, and Xuebin ZHANG

Abstract

In extreme value analysis, the impact of rounding in data, a form of quantization, on statistical inferences beyond point estimation has not been comprehensively studied. This paper addresses these challenges by considering rounded data as interval-censored. The maximum likelihood estimators of the model parameters tailored to account for interval censoring are asymptotically unbiased and efficient. Further, we adapt classic goodness-of-fit tests, such as the Anderson-Darling test, for rounded data based on the maximum likelihood estimator. The resulting tests have appropriate sizes and considerable power. One application of such tests is threshold selection for the peak over threshold approach in extreme value analysis. The efficacy of our estimation approach and the goodness-of-fit tests are demonstrated through a simulation study involving data rounded from generalized Pareto distributions. Applying this method to precipitation data from 18 stations in eastern Washington, an area with typically low precipitation and expecting a significant rounding effect, we observe narrower interval estimates of return levels.

KEYWORDS AND PHRASES: Discretized continuous distribution, Interval-censored, Quantized data, Threshold selection.

1. INTRODUCTION

Accurate extreme value modeling is essential in fields such as climate science and risk assessment, where return level estimation informs infrastructure planning and policy decisions. However, when data are rounded (i.e., quantized to a grid), standard statistical inference can become biased, leading to unreliable conclusions [12, 13]. One especially delicate area in the automated threshold selection procedure [4], where a sequence of goodness-of-fit tests are used to assess whether exceedances above a candidate threshold follow a generalized Pareto distribution (GPD). The procedure identifies the lowest threshold at which the GPD is statistically acceptable while controlling the False Discovery Rate (FDR) [16]. Because the selected threshold directly affects parameter estimation and return-level predictions, any distortion in the testing step can propagate into serious inferential error.

A practical workaround is jittering, a multiple-imputation method. In each imputation, every rounded value is perturbed within its rounding interval, the resulting data is analyzed as continuous data, and the results are pooled across imputations. This method is easy to implement and can work adequately when rounding is minor. However, it has drawbacks. It may induce bias in point estimates (as our simulations show). Its interval estimates depend on combining within- and between-imputation variances [28], which many users may not employ. And for goodness-of-fit tests, aggregating statistics across imputations (e.g., via medians) has been shown not to reliably reproduce the null distribution [4, Supplement, Figure S5]. Therefore, while jittering is a convenient workaround, it does not fully resolve the distortions caused by quantization, particularly for threshold choice and tail inference.

The impact of data rounding on extreme value analysis has been widely studied. Ignoring it leads to distorted sizes of goodness-of-fit tests [12] and poor performance in parameter estimation [13]. Solutions treating rounded data as interval-censored have emerged in applications dealing with rounded durations and count-based extreme value problems [24, 27]. Both studies use a discretized GPD, obtained by rounding a continuous GPD [18] to integers. One applied to dry-spell analysis in days [24], while the other modeled hospital congestion events based on daily emergency room visits [27]. While these methods leverage interval censoring to account for rounding effects, they do not generalize to continuous variables rounded at finer levels, such as precipitation recorded to the nearest 0.1 or 0.01 inches. No study has systematically examined the estimation of extreme precipitation under these rounding levels, and existing approaches provide no statistically valid goodness-of-fit test that maintains correct size under rounding.

Statistical methods for inference under rounding have been widely studied [17, 29]. A common assumption is that rounding errors follow a uniform distribution over a symmetric interval with length equal to the rounding precision and are independent of both the rounded and true values. Early work recognized that rounded data require a distinct likelihood function, treating them as interval-censored [20], and

*Corresponding author.

later research developed likelihood-based estimation methods for specific distributions [15, 33]. For serially dependent data, where the full likelihood is intractable, composite likelihood approaches have been introduced for quantized time series models [5], with consistency and asymptotic normality established [36]. Despite these advances, the adoption of statistical methods for rounded data in extreme value modeling remains limited among practitioners, with parameter estimation and goodness-of-fit testing under rounding receiving little attention until recently. While prior work examined how symmetric and asymmetric rounding affect estimation accuracy [29], systematic investigations into valid goodness-of-fit tests and their performance remain lacking, particularly for threshold-based extreme value analysis.

We provide a well-tested toolset for parameter estimation and goodness-of-fit testing under rounding error. The parameters are estimated by maximizing the true likelihood of the observed rounded data, constructed using interval censoring, which coincides with the continuous-data likelihood when no rounding error is present. For goodness-of-fit tests, we treat rounded data as discrete and adapt standard tests originally designed for continuous data, with p-values obtained through parametric bootstrap. This work makes three key contributions. First, we examine bias correction in parameter estimation under realistic rounding levels in precipitation data. Second, we study valid goodness-of-fit tests under rounding, demonstrating that tests adapted from continuous data, particularly the Anderson–Darling test, have higher power. Third, we show that applying a valid goodness-of-fit test directly impacts automated threshold selection, leading to different results in extreme value analysis. In extensive simulations, the MLEs of the GPD parameters appeared unbiased, goodness-of-fit tests held their sizes, and the adapted tests had substantial power. Applying this method to threshold selection in the POT framework for 18 eastern Washington stations resulted in different selection outcomes, leading to large differences in return level estimation.

The rest of the paper is organized as follows. Section 2 introduces the problem setup and presents estimation based on the true likelihood. Section 3 adapts classic goodness-of-fit tests for the GPD to handle rounded data. A simulation study in Section 4 evaluates the performance of the MLE and the goodness-of-fit tests. Section 5 applies the methods to automated threshold selection in extreme value analysis for precipitation at 18 sites in eastern Washington State. Section 6 concludes with a discussion.

2. LIKELIHOOD ESTIMATION

In extreme value theory, threshold exceedances are asymptotically modeled by GPD under broad conditions. This fundamental result [26] provides the theoretical foundation for using the GPD in threshold-based extreme value modeling. Without loss of generality, consider a GPD with location zero and distribution function

$$F(x;\sigma,\xi)=\begin{cases}1-\left[1+\frac{\xi x}{\sigma}\right]^{-1/\xi}, & \xi\neq 0;\\ 1-\exp\left(-\frac{x}{\sigma}\right), & \xi=0,\end{cases} \tag{2.1}$$

where $\sigma\in[0,\infty)$ and $\xi\in(-\infty,\infty)$ are scale and shape parameters, respectively. The support of the distribution is $x\geq 0$ if $\xi\geq 0$ and $0\leq x\leq -\sigma/\xi$ if $\xi<0$. Random samples $\mathbf{X}=(X_1,\ldots,X_n)$ of size n are drawn from $F(\cdot;\sigma,\xi)$. Without rounding, we would observe $\mathbf{x}=(x_1,\ldots,x_n)$. 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. The observed rounded sample $\mathbf{x}^*=(x_1^*,\ldots,x_n^*)$ satisfies $x_i^*=\delta\lceil x_i/\delta-0.5\rceil$ for $i=1,2,\ldots,n$, 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 $\mathbf{x}^*$.

When data are subject to rounding, the observed values follow a discretized version of the underlying continuous distribution. In the context of extreme value analysis, threshold exceedances of unrounded data are modeled by GPD, and the same principle applies when data are quantized. That is, the rounded data should follow a discretized GPD [24, 27, 18]. This forms the foundation of our likelihood construction. However, naive approaches that ignore this fact and treat rounded data as continuous can lead to substantial bias in parameter estimation [13]. These works highlight the impact of rounding but do not provide solutions, leaving the validity of statistical inference unresolved. Our approach explicitly incorporates discretization into the likelihood function, ensuring that estimation and goodness-of-fit tests remain valid under rounding effects.

The contribution of observation x_i^* to the likelihood is

$$\begin{aligned}&\Pr(X_i\in[(x_i^*-\delta/2)^+,x_i^*+\delta/2))\\ &=F\left(x_i^*+\frac{\delta}{2}\right)-F\left[\left(x_i^*-\frac{\delta}{2}\right)^+\right],\end{aligned} \tag{2.2}$$

where $t^+=\max(0,t)$ and $[(x_i^*-\delta/2)^+,x_i^*+\delta/2)$ is a half-open interval. 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 $F(x_i^*+\delta/2)=1$ in this case. The loglikelihood function is thus

$$\ell^*(\sigma,\xi;\mathbf{x}^*)=\sum_{i=1}^{n}\log\left\{F\left(x_i^*+\frac{\delta}{2}\right)-F\left[\left(x_i^*-\frac{\delta}{2}\right)^+\right]\right\}. \tag{2.3}$$

The maximizer of the loglikelihood (2.3), $(\hat\sigma,\hat\xi)$ is the MLE of (σ,ξ). As long as $\xi>-0.5$, the MLE is asymptotically unbiased and normally distributed [10], with variance being the inverse of the Fisher information matrix [31].

The likelihood contribution in (2.2) 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 (2.2) divided by δ converges to $f(x_i^*;\sigma,\xi)$ for $x_i^* > 0$ and to $f(0;\sigma,\xi)/2$ for $x_i^* = 0$. The resulting estimator, hereafter referred to as MLE-N, is similar to the true MLE only when δ is small relative to the scale of the distribution. As will be demonstrated in the simulation study, when δ increases, the bias of the MLE-N 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 $F(\cdot;\sigma,\xi)$ with 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 $t \in \{0,\delta,2\delta,\ldots\}$; 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

$$H_0 : \mathbf{x}^* \text{ is a random sample from distribution function } F_\delta(\cdot|\sigma,\xi) \text{ for some } (\sigma,\xi). \tag{3.1}$$

That is, we are testing that the data come from a discretized version of a GPD distribution with unknown parameters.

3.1 Chi-Squared Test

The first test to be considered is the chi-squared (CS) test [25], 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 MLE-N.

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 [14], 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} [7]. We approximate the p-value of the observed statistic through parametric bootstrap, where each bootstrap sample is generated from the fitted distribution with parameters $(\hat\sigma,\hat\xi)$ [32, Chapter 21]. This approach leverages distributional assumptions to create resamples, allowing for more efficient inference in cases where the assumption is appropriate and potentially improving accuracy for small samples or structured problems. See Appendix A for detailed steps.

In our numerical studies, the bins were chosen so that the fitted relative frequency for each bin is about 0.1.

3.2 Tests Adapted from the Continuous Case

The Kolmogorov–Smirnov (KS) [21, 30], Cramér–von Mieses (CvM) [11, 35], and Anderson–Darling (AD) [1] 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 MLE-N. 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 [9, 8]. When the hypothesized discrete distribution contains no unknown parameters, these tests have been adapted [2] and implemented in popular software packages [3]. 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; see Appendix A.

Since the observed data are rounded, the fitted distribution is discrete with probability mass function support points spaced at intervals of δ. Consequently, the test statistics depend on δ through the fitted distribution. However, δ does not explicitly appear in their expressions, as they are computed based on the empirical distribution of the discretized data. The limiting distributions of the test statistics, which are known for continuous data, are affected by the discrete nature of the fitted distribution. When δ is small relative to the data scale, the discretized empirical process closely approximates the continuous one, making the limiting distributions of the test statistics nearly the same as in the continuous case. However, for larger δ, deviations from the continuous limiting distributions may become more pronounced. The impact of δ on the empirical size and power of the tests is evaluated in the simulation study.

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

4. NUMERICAL STUDY

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 $\sigma \in \{0.3, 3\}$ and $\xi \in \{-0.1, 0, 0.1\}$. These settings were chosen to cover the estimated values of σ and ξ from the Eastern Washington State dataset, which are approximately 0.3 and 0.1, respectively. The rounding level δ was set to be 0, 0.01 to 0.1, where $\delta = 0$ meant no rounding, and $\delta \in \{0.01, 0.1\}$ match those the observed rounding levels in the dataset. In total, we had $2 \times 3 \times 3 = 18$ settings. For each setting, we generated 1000 datasets with sample size $n = 500$. The parametric bootstrap always has 1000 bootstrap replicates.

For comparison, the MLE-N obtained with the rounding ignored was also applied. 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 maximum likelihood estimates for both methods were obtained using the `Optim()` function in `Julia`, employing the default Nelder-Mead algorithm. [23, 19]. The MLE-N used the moment estimation as the initial value while MLE-IC used the MLE-N estimation as the initial value. The estimation for all the datasets converged. As shown in Figure S5 (Q-Q plots) in Supplementary Materials of [4], although the median AD statistic's quantiles of jittering are much closer than no adjustment, they still deviated significantly from the 45-degree line. Since the jittering method did not have a good fit, we only include this method in the point estimation.

4.1 Estimation

Figure 1 presents boxplots comparing the estimators from MLE-N, MLE-IC, and jittering methods for different rounding levels (δ), scaling parameters (σ), and location parameters (ξ). Notably, MLE-N and jittering methods exhibit bias when δ/σ is large. For example, in the case of $\sigma = 0.3$ and $\delta = 0.1$, both $\hat{\sigma}$ and $\hat{\xi}$ from MLE-N and jittering methods show greater bias compared to MLE-IC. Furthermore, the magnitude of this bias in MLE-N and jittering methods systematically increases with increasing δ. These findings align with previous studies [13]. As expected from its asymptotic properties, MLE-IC method exhibits negligible bias across all tested δ, σ, and ξ values. This observation aligns with recent literature [24]. The supplementary materials (Figure S-1) provide an additional more extreme case with $(\sigma, \xi) = (0.5, 0.2)$, and $\delta = 0.5$. In this more extreme scenario, the bias of MLE-N and jittering methods is even more pronounced, which further confirms the robustness of MLE-IC method.

We can also observe the mean squared errors (RMSE) in Figure 1. The pattern of the RMSE of the MLE-N and jittering methods 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 errors of the MLE-N and jittering methods are dominated by their biases. Conversely, the MLE-IC's RMSE is dominated by variance, which exhibits a minimal increase with increasing δ. This is intuitive since higher rounding levels imply less information, but the increase in variance here is negligible compared to MLE-N.

We also investigated the estimated and empirical standard errors of MLE-IC method, as shown in Figure 2. For all settings, the average estimated standard errors of MLE-IC closely matched the empirical standard errors. This suggests that the uncertainty associated with MLE-IC, critical for statistical inference, can be accurately estimated by inverting the Fisher information matrix for the sample size $n = 500$ were used in this study.

4.2 Goodness-of-Fit Test

We assessed the size and the power of the goodness-of-fit test. The size of a test is the maximum probability of rejecting the null hypothesis when it is true. The power of a test is the probability of rejecting the null hypothesis when it is false. The 4 goodness-of-fit test methods presented in Section 3 (CS, KS, CvM, and AD) were performed on each setting in Section 4.1. For all four 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.

4.2.1 Size

Figure 3 and 4 show the Q-Q plots of the 1000 p-values from tests for the settings with $(\sigma, \xi) = (0.3, 0.1)$ and $\delta \in \{0, 0.01, 0.1\}$. The results from other settings convey the same messages and, hence, are not 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 $(0, 1)$ and, consequently, that the Q-Q plots of the p-values to be around the 45 degree line in the unit square. A large deviation from the 45 degree line means that the null-distribution of the test statistics being used in the test is invalid and that the test would not hold its size. The expected Q-Q plot pattern is observed for four tests — CS, KS, CvM, and AD — when the fitted distribution were evaluated with MLE-IC. This is true regardless of the rounding level δ. For the MLE-N, when there is no rounding or rounding level is 0.01, the Q-Q plots are still around the 45 degree line. However, when the rounding level is 0.1, the Q-Q plots from MLE-N have a large deviation from the 45 degree line.

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. In particular, we generated data from distributions by truncating a hybrid uniform-GPD (HUGPD), which is uniform below a threshold μ with probability p and follows a GPD with distribution $F(\cdot; \sigma, \xi)$ above μ. The density function of this distribution is given in Equation (4.1)

$$g(x; \mu, \sigma, \xi) = \begin{cases} \frac{p}{\mu}, & 0 < x \leq \mu; \\ (1-p) f(x-\mu; \sigma, \xi), & x > \mu, \end{cases} \tag{4.1}$$

where $f(\cdot; \sigma, \xi)$ is the density of a GPD with distribution function $F(\cdot; \sigma, \xi)$ in Equation (2.1). To ensure continuity

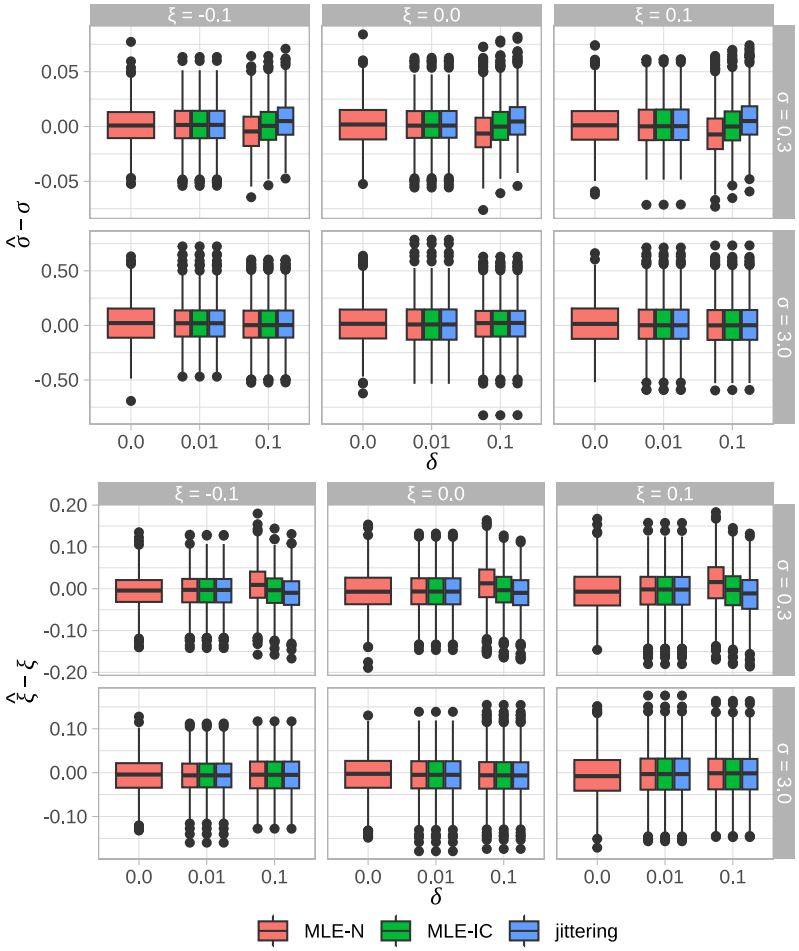

Figure 1: Boxplots of MLE-N, MLE-IC, and jittering methods. The upper panels show the estimated scale parameters $\hat{\sigma}$ and the lower panels show the estimated shape parameters $\hat{\xi}$.

at μ, only one of σ and ξ can be free for given (μ, p); we set $\sigma = p\mu/(1-p)$. The density of a truncated HUGPD with truncation point a is

$$g(x|x > a) = \frac{g(x) \cdot I(\{x > a\})}{1 - G(a)}, \tag{4.2}$$

where $G(x)$ is the distribution function of the HUGPD and $I(\cdot)$ is an indicator function.

In our simulation study, we set $\mu = 0.3$, $\xi = 0.1$, and $p \in \{0.8, 0.5, 0.2\}$, with corresponding $\sigma \in \{0.075, 0.3, 1.2\}$. See Figure 5 for the density functions of the three settings. The truncation point a were selected from 0.3 to 0 with step 0.05. As the truncation point decreases, the distribution shows more and more deviation from a GPD, and the power of any reasonably good test should increase. The sample size was set as 250 or 500 and the rounding level was set to be $\delta \in \{0, 0.01, 0.1\}$. 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 6 shows the empirical power of $g(x|x > a)$ for four statistical tests—AD, CvM, KS, and CS—applied to the HUGPD, with $\mu = 0.3$, $\sigma = 0.3$, $\xi = 0.1$, and $p = 0.5$. The AD and CvM tests outperform the others, with the KS test following and the CS test showing the least power. As expected, for a constant sample size, all tests' power increases when the truncation point moves from 0.3 to 0, indicating a greater deviation from the GPD. At a truncation point of 0.3, the powers of all tests approximate 0.05, suggesting that they maintain their nominal size. Particularly at a sample size of 500 and a truncation point of 0, the AD and CvM tests reach almost 100% power across all rounding levels, whereas the KS test's power drops to about 60% at a rounding level of 0.1, and the CS test's power remains low for all rounding levels. The power of all tests decreases as δ increases with a fixed truncation point and a sample size of 500, reflecting the expected data loss with higher δ. However, the decrease in power from $\delta = 0$ to $\delta \in 0.01, 0.1$ is minimal for the AD and CvM tests. Moreover, increasing the sample size from 250 to 500 enhances the power for all tests. In conclusion, the AD and CvM tests demonstrate superior performance in the settings of this study. Two additional

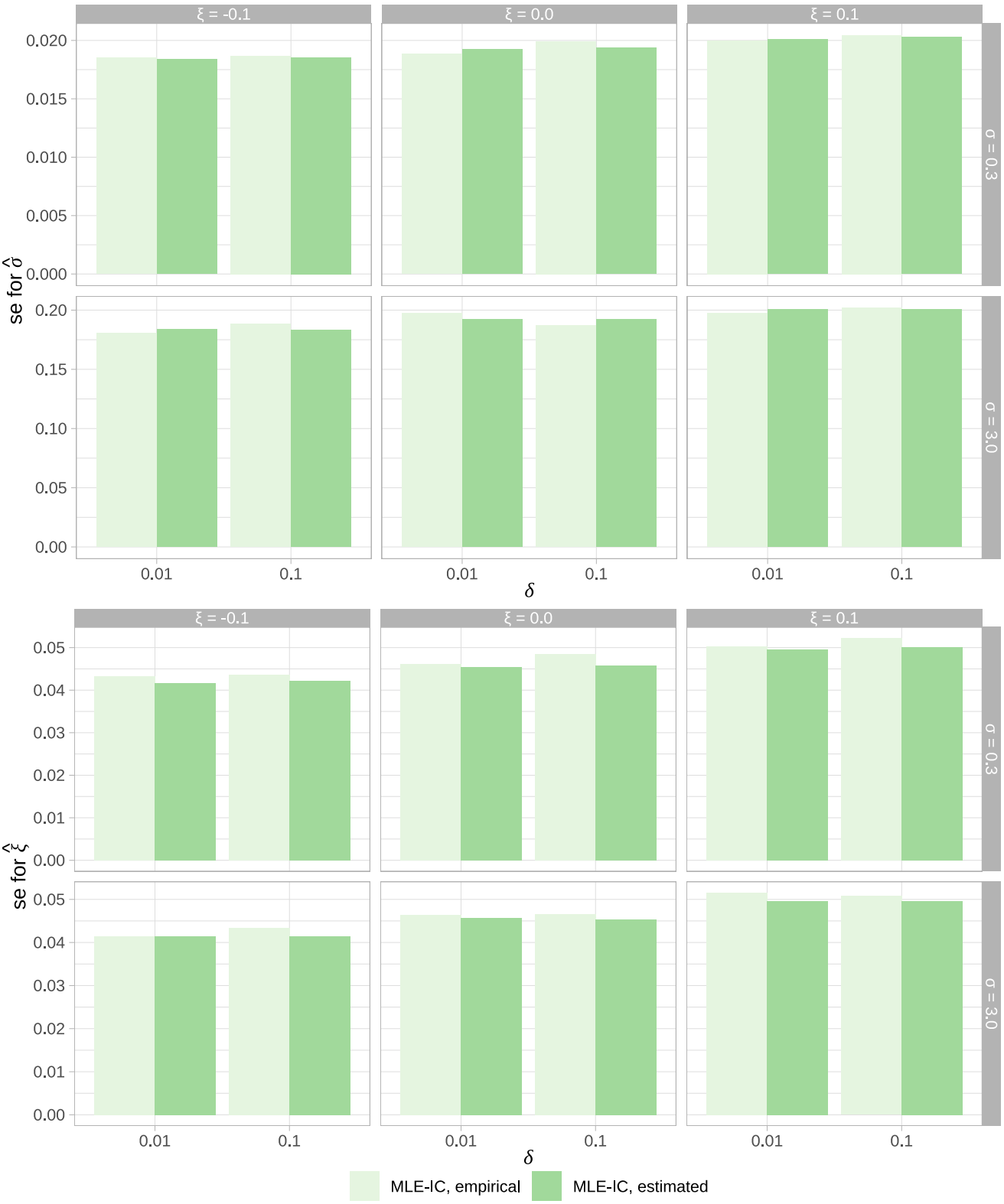

Figure 2: The estimated and empirical standard error of MLE-IC. The upper panels show the standard error of estimated scale parameter σ and the lower panels show the standard error of estimated shape parameter ξ.

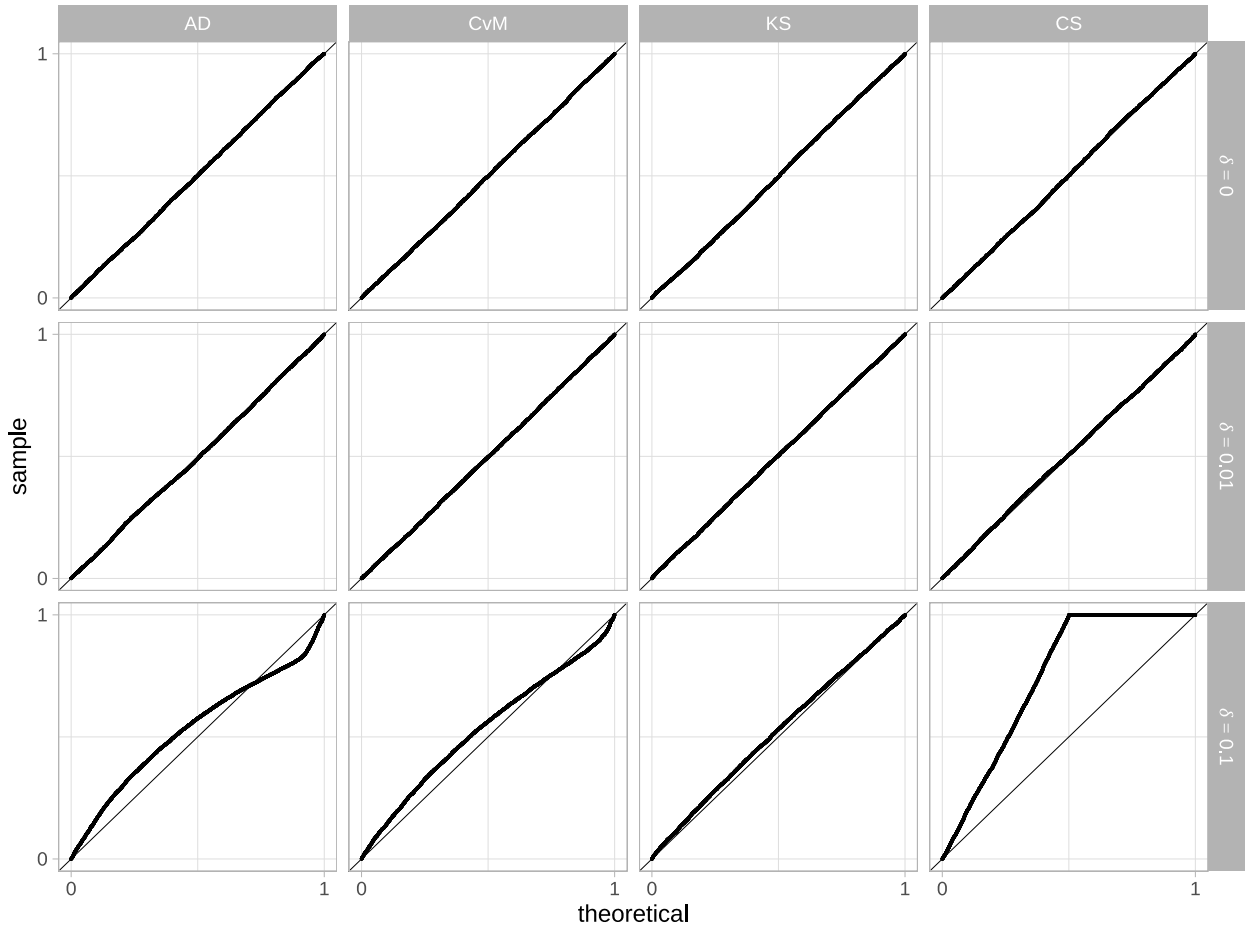

Figure 3: Q-Q plots of the p-values from 1000 replicates by the MLE-N method. Samples were generated by GPD with $(\sigma, \xi) = (0.3, 0.1)$, and no rounding (first row), $\delta = 0.01$ (second row), and $\delta = 0.1$ (third row).

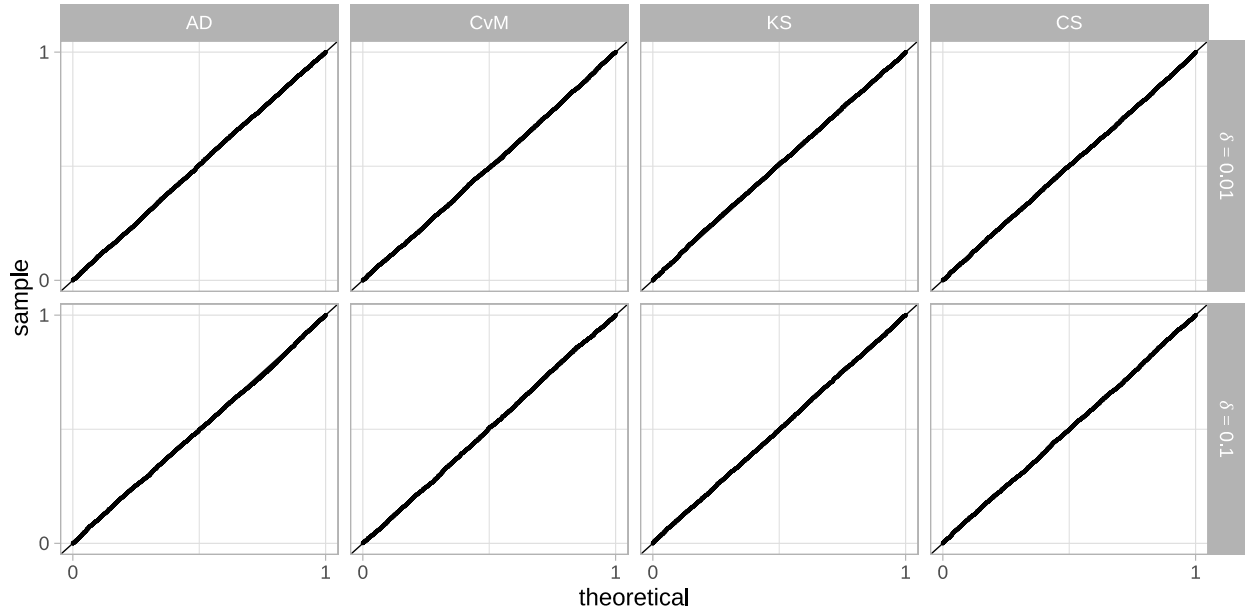

Figure 4: Q-Q plots of the p-values from 1000 replicates by the MLE-IC method. Samples were generated by GPD with $(\sigma, \xi) = (0.3, 0.1)$, and $\delta = 0.01$ (first row), and $\delta = 0.1$ (second row).

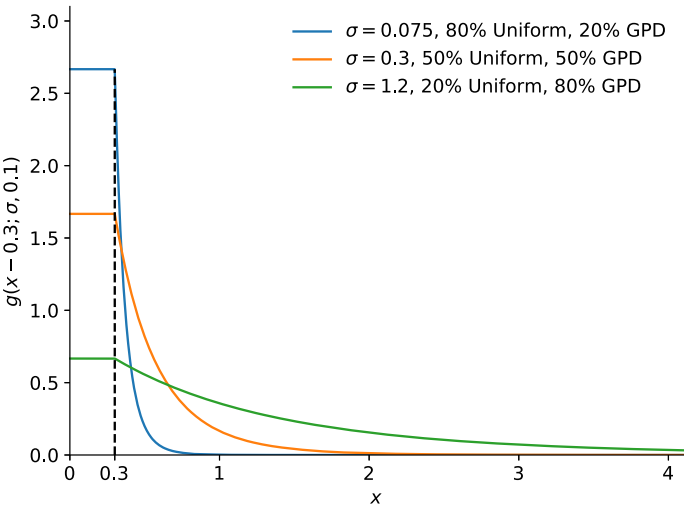

Figure 5: The probability density function of HUGPD $g(x-\mu;\sigma,\xi)$, where $\mu = 0.3$, $\sigma \in \{0.075, 0.3, 1.2\}$, $\xi = 0.1$, and $p \in \{0.8, 0.5, 0.2\}$.

plots from other settings are shown in Figure S-2 and S-3 in the Supplementary Material, where the GPD portion is the upper 20% and 80% of the mixed distribution. The AD and CvM tests consistently display the strongest sensitivity to deviations, while KS and CS tests show lower power. Doubling the sample size from 250 to 500 markedly increases power, confirming that the ranking of test strengths remains stable under more extreme scenarios.

5. APPLICATION TO ANNUAL MAXIMUM OF DAILY PRECIPITATIONS

The proposed method was applied to model daily precipitation data from 18 monitoring stations in eastern Washington State, covering the period from 1969 to 2018. This region is significantly drier (approximately 80–85% of days with no measurable precipitation) than the western part of the state due to the "rain shadow" effect of the Cascade Mountains. Since the simulation study suggested that the difference between MLE-IC and MLE-N is more pronounced in areas with less precipitation, this dataset provides an ideal setting to evaluate the impact of properly accounting for rounding. Daily precipitation data were obtained from the Global Historical Climatology Network [22]. Given that most precipitation occurs in winter, we restricted the analysis to winter months (November through March), resulting in a total of 7512 winter days from 1969 to 2018. For each site, we tested 15 candidate thresholds, taking the 70th to 98th percentiles in increments of 2 percent. Although zero precipitation values are included in the dataset, the lowest candidate threshold at each station is greater than zero. Precipitation amounts are recorded to the nearest hundredth of an inch, meaning $\delta = 0.01$. Missing data were ignored in the analysis. See Section 2 in the Supplementary Material for additional details of the analysis.

The automated threshold selection procedure [4] relies critically on the goodness-of-fit test for the GPD at each candidate threshold. It uses the ForwardStop procedure [16] to control the FDR in the sequential testing of ordered null hypotheses; see steps in Appendix B. However, rounding effects were not accounted for in the test statistic, which our method resolves. We focus on the AD test, which demonstrated the highest power in the simulation study. The results of the sequential AD tests lead to three possible outcomes: (1) no threshold is selected by either MLE-IC or MLE-N; (2) a threshold is selected by MLE-IC but not by MLE-N; or (3) a threshold is selected by both methods, but MLE-IC chooses a lower threshold than MLE-N. The second and third cases highlight the advantage of MLE-IC in selecting thresholds with more exceedances, leading to higher efficiency in statistical inference.

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. Table S-2 in the Supplementary Material summarizes the candidate thresholds and the corresponding number of exceedances at the two sites. Figure S-5 in the Supplementary Material shows the p-values at the 15 candidate thresholds at the two sites before and after the ForwardStop adjustment using the AD test [4]. At Chewelah, the tests based on MLE-N selects threshold 1.75 with 135 exceedances; the tests based on MLE-IC selects 0.25, which is the 80th percentile, 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 7 shows the parameter estimates $\hat{\sigma}$ and $\hat{\xi}$ and their 95% confidence intervals for the two sites if a threshold was selected. The 95% confidence intervals of $\hat{\sigma}$ and $\hat{\xi}$ are constructed as $\hat{\sigma} \pm z_{0.25}\widehat{\mathrm{SD}(\hat{\sigma})}$ and $\hat{\xi} \pm z_{0.025}\widehat{\mathrm{SD}(\hat{\xi})}$, respectively, where $z_{0.025}$ is the 97.5% upper quantile of the standard normal distribution and $\widehat{\mathrm{SD}(\cdot)}$ is the standard error of estimate. 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. Figure S-7 in the Supplementary Material presents the estimated 25-, 50-, 100-, and 200-year return levels along with 95% confidence intervals constructed from profile likelihood for the

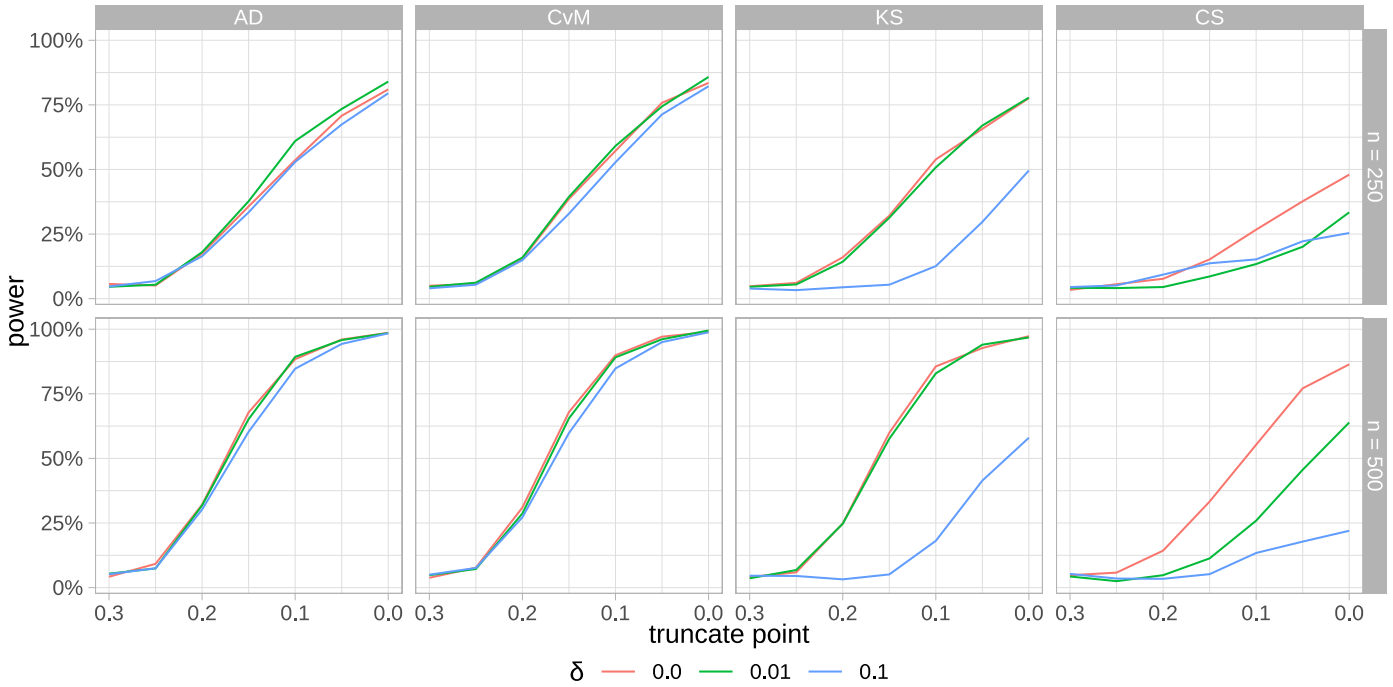

Figure 6: Power of $g(x|x > a)$ for four tests applied to the HUGPD, with $\mu = 0.3$, $\sigma = 0.3$, $\xi = 0.1$, and $p = 0.5$. Sample size is 250 and 500.

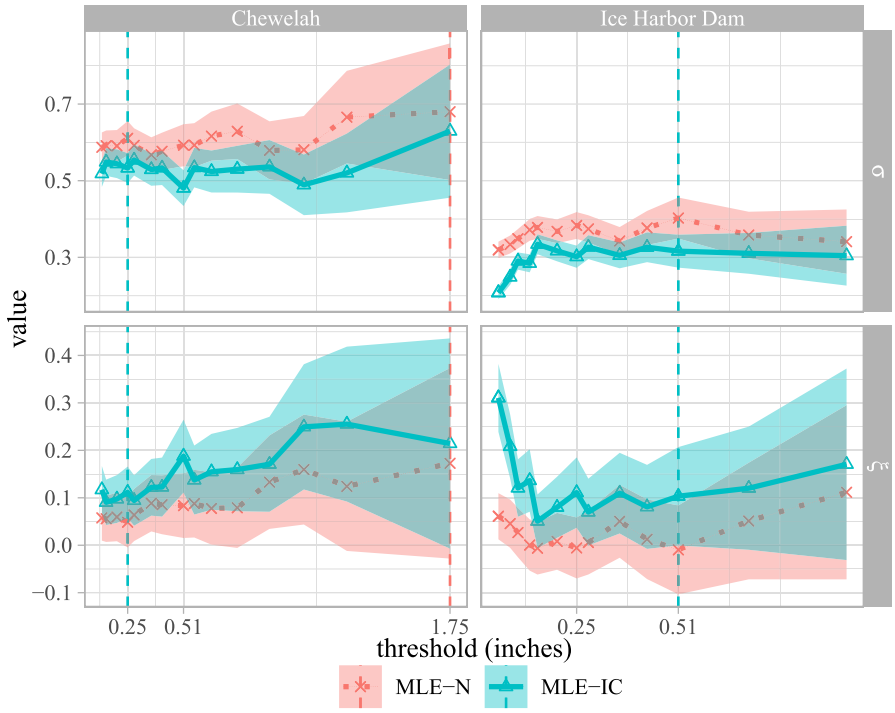

Figure 7: GPD parameter estimates and 95% confidence intervals. The MLE-N and MLE-IC are obviously different in both point estimates and uncertainty.

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. In the Supplementary Material, the time series plots (Figure S-4 in the Supplementary Material) with thresholds display the precipitations of the two stations over time, and the Q-Q plots (Figure S-6 in the Supplementary Material) show that MLE-IC has a better fit than MLE-N.

We performed an analysis of the return levels for each of the 18 sites; the selected thresholds and the corresponding number of exceedances are summarized in Table S-1 in the Supplementary Material. Among all 18 stations in eastern Washington, 9 stations had thresholds selected via MLE-IC, but only 2 stations 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 [4] 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 [4], 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 rounding error in extreme value analyses [12, 13] 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 broad applications in extreme value analyses of precipitation or temperature, potentially leading to significantly different results than those obtained otherwise. When the generalized extreme value (GEV) distribution is used, or when some parameters incorporate covariates, the interval censoring framework can be applied as well. This extension may be particularly relevant in settings where block maxima are subject to rounding.

The correction of the method to the naive approach depends on the rounding level relative to the scale parameter. A change in the measurement unit, for example, from cm to inch, would result in the same scaling effect on the estimated scale parameter of GPD accordingly, but the estimated shape parameter and the p-value of the goodness-of-fit should remain the same. The method requires that the rounding level δ is known and can handle datasets with multiple rounding levels, such as cases where later data are recorded with higher precision. Although the AD test for quantized data has the highest power, it is much more computing intensive than the CS test, especially when δ is small relative to σ and $\xi > 0$, as this increases the number of discrete support points. A faster alternative would be of interest. Additionally, in real data applications, the annual maximum of daily precipitations may have serial dependence, which could affect the accuracy of the proposed methods. Addressing serial dependence merits further research.

The impact of correcting the bias with MLE-IC is greatest at locations with lower precipitation. Our illustration focused on 18 eastern Washington stations, which are known to be much drier than those to the west of the mountains. In batch studies [4], where a large number of individual sites are analyzed one by one, we expect more pronounced differences if MLE-IC were used instead of MLE-N, in terms of the number of stations with a threshold selected, the number of exceedances, and the resulting point and interval estimates of various return levels. Since threshold selection is beyond the scope of this paper, we did not study its uncertainty. Bayesian methods could provide a natural approach for handling this [6, 34].

SUPPLEMENTARY MATERIAL

The Supplementary Material presented additional simulation results confirming MLE-IC's robustness under extreme rounding conditions and extended power comparisons. It also provided supplementary precipitation data analysis across 18 eastern Washington stations, showing that MLE-IC consistently selected lower thresholds and yielded better model fit and return-level estimates than MLE-N.

APPENDIX A. GOODNESS-OF-FIT TEST WITH PARAMETRIC BOOTSTRAP

The parametric bootstrap procedure for a goodness-of-fit test with testing statistics S that depends on some estimator of the parameters are summarized as follows.

1. Obtain estimates of GPD parameters $(\hat{\sigma}, \hat{\xi})$ from the data $\mathbf{x}^*$.
2. Calculate test statistic S based on $(\hat{\sigma}, \hat{\xi})$.
3. For $i \in \{1, \ldots, B\}$, where B is a large number representing the number of bootstrap replicates:
   (a) Generate a sample $\mathbf{x}^{(i)}$ of size n from a GPD with fitted parameter $(\hat{\sigma}, \hat{\xi})$.
   (b) Round $\mathbf{x}^{(i)}$ to $\mathbf{x}^{(i)*}$ with rounding level δ.

(c) Obtain GPD parameter $(\hat{\sigma}^{(i)}, \hat{\xi}^{(i)})$ based on bootstrap sample $\mathbf{x}^{(i)*}$.

(d) Calculate the test statistic $S^{(i)}$ based on $(\hat{\sigma}^{(i)}, \hat{\xi}^{(i)})$.

4. Approximate p-value by

$$\hat{p} = \frac{0.5 + \sum_{b=1}^{B} I(S^{(b)} > S)}{B+1},$$

where $I(\cdot)$ the indicator function.

APPENDIX B. FORWARDSTOP IN ORDERED HYPOTHESES TESTING

ForwardStop is used to control the false discovery rate at level α. It has demonstrated resilience to various data distributions, including those with correlated p-values, which is relevant to our threshold selection process[16]. Consider a sequence of ordered hypothesis $H_0^{(i)}$, $i = 1, \ldots, L$. They are ordered in the sense that if $H_0^{(i)}$ is rejected, then all $H_0^{(j)}$ for $j < i$ will be rejected. In the context of threshold selection with candidate thresholds $u_1 < \cdots < u_L$ [4], the sequence of null hypotheses are

$H_0^{(i)}$: the distribution of the observations exceeding u_i follow a GPD.

That is, if the u_i is rejected as a threshold, then any threshold lower than u_j would be rejected.

The detailed steps of calculating ForwardStop adjusted p-values are as follows.

1. Obtain the p-values $p_1, \ldots, p_L$ of the L hypotheses.
2. Arrange the p-values in reverse order $p'_1 = p_L, \ldots, p'_L = p_1$.
3. For $k = 1, \ldots, L$, calculate the corresponding $q'_1, \ldots, q'_L$ by $q_k = \frac{1}{k}\sum_{j=1}^{k} -\log(1 - p'_k)$.
4. Arrange the p-values $q'_1, \ldots, q'_L$ in reverse order to $q_1 = q'_L, \ldots, q_L = q'_1$, and then q_k is the corresponding ForwardStop adjusted p-value of p_k for $k = 1, \ldots, L$.

FUNDING

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

Accepted 3 February 2026

REFERENCES

[1] Anderson, T. W. and Darling, D. A. (1952). Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics* **23**(2) 193–212. https://doi.org/10.1214/aoms/1177729437. MR0050238

[2] Arnold, T. B. and Emerson, J. W. (2011). Nonparametric Goodness-of-fit Tests for Discrete Null Distributions. *R Journal* **3**(2) 34–39.

[3] Arnold, T. B. and Emerson, J. W. (2013). dgof: Discrete Goodness-of-Fit Tests. R package version 1.2. https://CRAN.R-project.org/package=dgof.

[4] Bader, B., Yan, J. and 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. https://doi.org/10.1214/17-AOAS1092. MR3773395

[5] Bai, Z., Zheng, S., Zhang, B. and Hu, G. (2009). Statistical Analysis for Rounded Data. *Journal of Statistical Planning and Inference* **139**(8) 2526–2542. https://doi.org/10.1016/j.jspi.2008.11.018. MR2523645

[6] Behrens, C. N., Lopes, H. F. and Gamerman, D. (2004). Bayesian Analysis of Extreme Events with Threshold Estimation. *Statistical Modelling* **4**(3) 227–244. https://doi.org/10.1191/1471082X04st075oa. MR2062102

[7] Chernoff, H. and 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. https://doi.org/10.1214/aoms/1177728726. MR0065109

[8] Choulakian, V., Lockhart, R. A. and Stephens, M. A. (1994). Cramér-von Mises Statistics for Discrete Distributions. *Canadian Journal of Statistics* **22**(1) 125–137. https://doi.org/10.2307/3315828. MR1271450

[9] Conover, W. J. (1972). A Kolmogorov Goodness-of-fit Test for Discontinuous Distributions. *Journal of the American Statistical Association* **67**(339) 591–596. MR0391375

[10] Cramér, H. (1946). A Contribution to The Theory of Statistical Estimation. *Scandinavian Actuarial Journal* **1946**(1) 85–94. https://doi.org/10.1080/03461238.1946.10419631. MR0017505

[11] Cramér, H. (1928). On the Composition of Elementary Errors. *Scandinavian Actuarial Journal* **1928**(1) 13–74.

[12] 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**(18) 1240–1251.

[13] 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**(10–12) 626–634.

[14] 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.

[15] Giesbrecht, F. and 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. MR0652563

[16] G'Sell, M. G., Wager, S., Chouldechova, A. and Tibshirani, R. (2016). Sequential Selection Procedures and False Discovery Rate Control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**(2) 423–444. https://doi.org/10.1111/rssb.12122. MR3454203

[17] Heitjan, D. F. (1989). Inference from Grouped Continuous Data: A Review. *Statistical Science* **4** 164–179.

[18] Hitz, A. S., Davis, R. A. and Samorodnitsky, G. (2024). Discrete Extremes. *Journal of Data Science* **22**(4) 524–536.

[19] Jalbert, J., Farmer, M., Gobeil, G. and Roy, P. (2024). Extremes.jl: Extreme Value Analysis in Julia. *Journal of Statistical Software* **109**(6) 1–35. https://doi.org/10.18637/jss.v109.i06.

[20] Kempthorne, O. (1966). Some Aspects of Experimental Inference. *Journal of the American Statistical Association* **61**(313) 11–34. MR0195211

[21] Kolmogorov, A. (1933). Sulla Determinazione Empirica di una Legge di Distribuzione. *Giornale dell'Istituto Italiano degli Attuari* **4** 83–91.

[22] Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E. and Houston, T. G. (2012). An Overview of the Global Historical Cli-

matology Network-daily Database. *Journal of Atmospheric and Oceanic Technology* **29**(7) 897–910.

[23] Mogensen, P. and Riseth, A. (2018). Optim: A Mathematical Optimization Package For Julia. *Journal of Open Source Software* **3**(24).

[24] Pasarić, Z. and Cindrić, K. (2019). Generalised Pareto Distribution: Impact of Rounding on Parameter Estimation. *Theoretical and Applied Climatology* **136**(1) 417–427.

[25] 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.

[26] Pickands, J. (1975). Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics* **3**(1) 119–131. MR0423667

[27] Ranjbar, S., Cantoni, E., Chavez-Demoulin, V., Marra, G., Radice, R. and Jaton-Ogay, K. (2022). Modelling the Extremes of Seasonal Viruses and Hospital Congestion: The Example of Flu in a Swiss Hospital. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **71**(4) 884–905. https://doi.org/10.1111/rssc.12559. MR4470824

[28] Rubin, D. B. (2004) *Multiple Imputation for Nonresponse in Surveys* **81**. John Wiley & Sons. MR2117498

[29] Schneeweiss, H., Komlos, J. and Ahmad, A. S. (2010). Symmetric and Asymmetric Rounding: A Review and Some New Results. *AStA Advances in Statistical Analysis* **94**(3) 247–271. https://doi.org/10.1007/s10182-010-0125-2. MR2733174

[30] Smirnov, N. (1939). On the Estimation of the Discrepancy Between Empirical Curves of Distribution for Two Independent Samples. *Bulletin Mathématique de l'Université de Moscou* **2** 3–14. MR0002062

[31] Smith, R. L. (1985). Maximum Likelihood Estimation in a Class of Nonregular Cases. *Biometrika* **72**(1) 67–90. https://doi.org/10.1093/biomet/72.1.67. MR0790201

[32] Tibshirani, R. J. and Efron, B. (1993) *An Introduction to the Bootstrap*. Chapman & Hall, New York. https://doi.org/10.1007/978-1-4899-4541-9. MR1270903

[33] Vardeman, S. B. and Lee, C. q. q. S. (2005). Likelihood-based Statistical Estimation from Quantized Data. *IEEE Transactions on Instrumentation and Measurement* **54**(1) 409–414.

[34] Villa, C. (2017). Bayesian Estimation of the Threshold of a Generalised Pareto Distribution for Heavy-tailed Observations. *Test* **26**(1) 95–118. https://doi.org/10.1007/s11749-016-0501-7. MR3613607

[35] von Mises, R. (1931) *Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und Theoretischen Physik*. Deuticke, Leipzig und Wien.

[36] Zhang, B., Liu, T. and Bai, Z. (2010). Analysis of Rounded Data from Dependent Sequences. *Annals of the Institute of Statistical Mathematics* **62**(6) 1143–1173. https://doi.org/10.1007/s10463-009-0224-6. MR2729157

Sai Ma. Department of Statistics, University of Connecticut, Storrs, CT, USA.
E-mail address: sai.ma@uconn.edu

Jun Yan. Department of Statistics, University of Connecticut, Storrs, CT, USA.
E-mail address: jun.yan@uconn.edu

Xuebin Zhang. Pacific Climate Impacts Consortium, University of Victoria, Victoria, BC, Canada.
E-mail address: xuebinzhang23@uvic.ca