

Some Parameter Estimators in the Generalized Pareto Model and their Inconsistency with Observed Data

F. Ashkar,[1] and C. N. Tatsambon[2]

[1]Department of Mathematics and Statistics, Université de Moncton, Moncton, N.B., Canada, E1A 3E9; PH (506) 858-4312; FAX (506) 858-4396; email: ashkarf@umoncton.ca
[2]Department of Mathematics and Statistics, Université de Moncton, Moncton, N.B., Canada, E1A 3E9; PH (506) 858-4298; FAX (506) 858-4396; email: ntakala@yahoo.fr

Abstract

The generalized Pareto distribution (GPD) is widely used in the frequency modeling of hydrological extremes. Statistical methods used to fit this model to data include the methods of maximum likelihood (ML), of moments (MM), of probability weighted moments (PWM), and of generalized probability weighted moments (GPWM). When the shape parameter of the GPD is positive, the sample space is a finite interval whose upper bound depends on the distribution parameters. The MM, PWM and GPWM methods may produce estimates of this upper bound that are inconsistent with the observed data. This inconsistency occurs when one or more sample observations exceed the estimated upper bound, thus making this estimated upper bound physically unjustifiable. In this paper we shed more light on this problem of inconsistency with the data and examine its consequences by using Monte Carlo simulation. We provide new guidelines for choosing between the ML, MM, PWM and GPWM methods for estimating GPD quantiles.

Introduction

The generalized Pareto distribution (GPD) is widely used in hydrological frequency analysis, particularly in the peaks-over-threshold (POT) approach for modeling hydrological extremes. It is applied in situations in which the exponential distribution might be appropriate, but robustness is desired against a heavier or a thinner tail than that of the exponential model.

Methods that are used to fit the GPD to hydrological data include the method of maximum likelihood (ML), the method of moments (MM), and the method of probability weighted moments (PWM). Also, the method of generalized probability weighted moments (GPWM) has been proposed by Rasmussen (2001) but is still not widely used.

A GPD random variable *X* has the probability density function

$$f(x;b,k) = \frac{1}{b}(1 - k\frac{x}{b})^{(1/k)-1} \quad k \neq 0$$

$$f\left(x;b\right) = \frac{1}{b}e^{-x/b} \quad k = 0$$

The range of X is $0 \leq x < \infty$ for $k < 0$ and $0 \leq x \leq b/k$ for $k > 0$. It is readily seen that b is a scale parameter and k is a shape parameter. It is also seen that when $k > 0$, the sample space of X has the upper bound (b/k), which depends on the distribution parameters.

It has recently been pointed out in the statistical literature (Dupuis, 1996), that certain estimation methods (e.g., MM, PWM) may produce estimates of this upper bound that are inconsistent with the observed data. This inconsistency occurs when one or more sample observations exceed the estimated upper bound. If this happens, it is obvious that the estimated upper bound, $\hat{b}/\hat{k}$, becomes physically unjustifiable.

This problem of "unfeasible parameter estimates", or "unfeasible estimated upper bound", has not yet been given the attention that it requires in the hydrological literature. This paper sheds more light on this problem and examines its consequences by use of Monte Carlo simulation.

Specific Objectives

Hosking and Wallis (1987) used simulation with the GPD model and found that the MM and PWM methods produce smaller bias and root mean square error (RMSE) than the ML method for samples of size less than 500. Their results have led to a wide use of MM and PWM methods with the GPD model in hydrological practice. However, these results need now to be reexamined by taking into account the problem of "unfeasible parameter estimates" that has just been mentioned. At the same time, the recommendations given by Rasmussen (2001), with regard to the GPWM method, will also have to be reexamined.

Our first objective will be to test how often an estimation method (ML, MM, PWM or GPWM) produces an estimate of the GPD upper bound $\hat{b}/\hat{k}$ that is inconsistent with the simulated data. Specifically, we will try to find out, using computer generated samples, how often do we have $\hat{k} > 0$ and one or more sample observations exceed $\hat{b}/\hat{k}$. In other words, we will try to estimate and analyse the "inconsistency rate" of the four estimation methods under investigation. As we will see later, the analysis of this inconsistency rate will make a second objective necessary, which is to reexamine the simulation results obtained by Hosking and Wallis (1987) (ML, MM and PWM methods), and by Rasmussen (2001) (GPWM method). However, before embarking on these two objectives, we will review some of the main results obtained by Hosking and Wallis (1987), and by Rasmussen (2001).

Overview of Some Earlier Results

1. Study by Hosking and Wallis (1987). As mentioned earlier, Hosking and Wallis (1987) compared estimators of GPD quantiles obtained by the ML, PWM and MM methods. They found MM quantile estimators to be preferable to other estimators for $k > 0$ or $k \approx 0$. Only for $k < -0.2$ did the PWM method give superior results to the MM method. These authors also reported poor results for the ML method for small to moderate sample sizes frequently encountered in hydrology. They recommended the ML method only for very large sample sizes, and only for $k > 0.2$.

Hosking and Wallis (1987) also encountered serious numerical problems with the ML estimation algorithm that they employed. Their algorithm produced very high failure rates, where a failure rate is defined as the percentage of times that an estimation algorithm fails to converge. For example, with $k = 0.4$ and sample sizes of 15 and 25, the failure rates obtained by Hosking and Wallis (1987) were in excess of 41% and 14% respectively.

A consequence of the high failure rates obtained by Hosking and Wallis (1987) with their ML estimation algorithm, is that the ML method was much less used in subsequent years in hydrology for fitting the generalized Pareto distribution. The ML method was particularly ignored in two important hydrological studies, one by Dupuis (1996), and the other by Rasmussen (2001).

2. Study by Rasmussen (2001). Rasmussen (2001) gave a practical decision rule on how to apply the GPWM method to estimate GPD quantiles x_T, for $T \geq 50$. For these upper tail quantiles, he found that: (1) the GPWM method systematically outperformed the (classical) PWM method; (2) the performance of the GPWM method relative to the PWM method improved with increasing k, and increasing T; (3) the PWM method performed better for $k < 0$ than for $k > 0$. All these conclusions were based on an analysis of the mean square errors of quantile estimates (equivalently, RMSEs).

In comparing the GPWM and MM methods for estimating GPD quantiles x_T, Rasmussen (2001) found that: (1) the GPWM method systematically outperformed the MM method for $T = 50$ and for $T = 100$; (2) for k values in the interval [-0.3, 0.3], which are the values most frequently encountered in practice, the difference between the GPWM and MM methods was small; (3) for larger T values (e.g. $T = 1000$), the MM method outperformed the GPWM method for k values in the range of practical interest.

Monte Carlo Simulations

After having reviewed the main results obtained by Hosking and Wallis (1987) and by Rasmussen (2001), we will now embark on the two objectives that we had set earlier.

First objective: Testing the inconsistency with simulated data. We will test how often each of the four estimation methods under investigation, produces an estimate of the GPD upper bound that is inconsistent with data that will be generated by computer.

Computer simulations were carried out with several sample sizes, n, and several shape parameter values, k, of the GPD population. The values $n = 15, 25, 50$ and 100 were chosen, along with k values from -0.4 to 0.4, in steps of 0.1. With no loss of generality, the scale parameter b was set equal to 1. For each combination of sample size, n, and shape parameter value, k, 1000 GPD samples were randomly generated.

The ML, MM and PWM methods were applied according to Sections 3.1, 3.2 and 3.3, respectively, of (Hosking and Wallis, 1987), whereas the GPWM method was applied as described in Section 5.2 of (Rasmussen, 2001). More specifically, Equations 18 (a, b), of (Rasmussen, 2001), were used as a practical decision rule on how to apply the GPWM method with any generated sample. See Section 5.2 of (Rasmussen, 2001) for more detail.

In the PWM and GPWM methods, the sample PWMs and GPWMs were estimated using the plotting position estimate, p_i, of $F(x)$, given by:

$$\hat{F}(x_{(i)}) = p_i = \frac{i - 0.35}{n}$$

where $F(x)$ is the cumulative distribution function (CDF) of X. In this expression, $x_{(i)}$ denotes the ith element of the ordered sample $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ .This expression for p_i is the one used both by Hosking and Wallis (1987) and by Rasmussen (2001).

Table 1 shows, for each (n, k) combination, the percentage of times that each method of estimation produced an estimate of the GPD upper bound that is inconsistent with the simulated data.

It is seen from Table 1 that for several (n, k) combinations, the MM, PWM and GPWM methods have a serious problem of inconsistency with the simulated data. As expected, this problem is more serious for $k > 0$ and becomes more and more severe as k increases from 0 to 0.4. It is seen that the “inconsistency rate” exhibited by any of the three methods just mentioned, can be as high as 15, 20 or even 30 %.

With the MM and PWM methods, the problem of high inconsistency rate persists with all the sample sizes that were considered. In fact, it can be seen that for k close to 0.4, the inconsistency rate increases with increasing sample size. A similar observation was made by Dupuis (1996). For $n = 15$, the problem of inconsistency exhibited by the GPWM method seems to be more serious than that exhibited by either the MM or the PWM methods. With $n = 25$, the problem with the GPWM method still persists, but becomes less serious for $n \geq 50$.

It is important to note, however, that the ML method does not share the problem of inconsistency with the simulated data that is displayed by the other three methods (Table 1). The ML estimation algorithm employed in the present study is one that was proposed by Choulakian and Stephens (2001). It is also important to report that with this algorithm, no numerical problems similar to the ones reported by Hosking and Wallis (1987) were encountered. In fact, the estimation algorithm that we employed, never failed to converge.

Second objective: Reexamination of earlier results. The problems of inconsistency with observed data of the kind demonstrated in the previous section have up till now been largely overlooked in the hydrological literature. But due to the seriousness of this problem, some of the comparisons between the ML, MM, PWM and GPWM methods that have previously been reported in the hydrological literature, will have to be reexamined. In particular, we need to re-assess the results obtained by Hosking and Wallis (1987), and by Rasmussen (2001).

Table 1 Percentage of times that each method of estimation produced an estimate of the GPD upper bound that is inconsistent with the simulated data. We have called these percentages "inconsistency rates". A shaded area corresponds to an inconsistency rate greater than 5%.

n	Method	*k*								
		-0.4	-0.3	-0.2	-0.1	0	0.1	0.2	0.3	0.4
15	ML	0	0	0	0	0	0	0	0	0
	MM	0.7	0.4	1.1	1.5	2.9	4.8	7.3	10.1	13.5
	PWM	1.2	1.6	2.7	3.3	5.8	8.1	11.2	13.9	16.7
	GPWM$^{(*)}$	3.6	4.4	4.9	7.9	10.4	14.1	17.5	27.6	34.5
25	ML	0	0	0	0	0	0	0	0	0
	MM	0	0	0.5	0.5	1.5	4.2	6.4	12.5	15.1
	PWM	0.5	1.1	1	2.1	4.2	9.5	11.2	16.8	18.8
	GPWM	0.1	0.3	0.4	0.5	0.8	2.4	4.4	7.9	12.3
50	ML	0	0	0	0	0	0	0	0	0
	MM	0	0	0	0.3	0.6	1.6	4.4	11.1	15
	PWM	0	0	0.5	0.9	2.6	5.8	9.5	16.4	19.9
	GPWM	0	0	0	0.1	0	0.2	0.5	1.3	3.2
100	ML	0	0	0	0	0	0	0	0	0
	MM	0	0	0	0	0.2	0.6	3.3	10.8	15.9
	PWM	0	0	0	0.4	1.6	2.8	8.4	17	21.8
	GPWM	0	0	0	0	0	0.1	0.1	0.8	2.3

$^{(*)}$Inconsistency rates reported for the GPWM method are obtained by applying Equations 18 (a, b) of Rasmussen (2001); they are essentially the same whatever value of T is inserted into these equations.

This re-assessment will be based on the same computer simulations that were outlined earlier. With each GPD simulated sample, parameter estimates $\hat{k}$ and $\hat{b}$ were obtained, as well as quantile estimates, $\hat{x}_T$ for return periods $T = 10$, 50, 100 and 200. The biases and RMSEs for $\hat{k}$, $\hat{b}$ and $\hat{x}_T$ were then calculated. However, only results pertaining to $\hat{x}_T$ will be reported herein, since they are the ones of most interest in hydrological practice.

Table 2 presents the RMSEs for $\hat{x}_T$, $T = 100$, for different k values of the GPD population, and different sample sizes, n.

Table 2 RMSEs of GPD quantile estimates corresponding to a return period T = 100. For each (k, n) combination, a shaded cell indicates the lowest RMSE that was obtained. Cells marked with "\" correspond to an inconsistency rate above 10%; those marked with "/"correspond to an inconsistency rate of between 5 and 10 %. Exact inconsistency rates can be read from Table 1.

n	Method	*k*								
		-0.4	-0.3	-0.2	-0.1	0	0.1	0.2	0.3	0.4
15	ML	1.15	1.05	0.81	0.64	0.48	0.42	0.34	0.28	0.23
	MM	0.72	0.53	0.47	0.4	0.35	0.33	0.29	0.25	0.23
	PWM	0.72	0.63	0.57	0.5	0.43	0.4	0.35	0.31	0.29
	GPWM	0.53	0.51	0.48	0.43	0.39	0.37	0.33	0.29	0.25
25	ML	0.81	0.68	0.58	0.47	0.36	0.3	0.25	0.2	0.16
	MM	0.56	0.42	0.38	0.34	0.29	0.26	0.22	0.2	0.18
	PWM	0.59	0.5	0.46	0.4	0.34	0.31	0.27	0.25	0.23
	GPWM	0.45	0.41	0.38	0.35	0.31	0.28	0.25	0.21	0.18
50	ML	0.52	0.47	0.42	0.31	0.26	0.2	0.16	0.13	0.1
	MM	0.4	0.32	0.3	0.25	0.21	0.18	0.15	0.14	0.13
	PWM	0.44	0.37	0.35	0.28	0.25	0.22	0.19	0.17	0.16
	GPWM	0.36	0.32	0.3	0.26	0.23	0.19	0.16	0.14	0.11
100	ML	0.37	0.31	0.27	0.22	0.18	0.14	0.11	0.09	0.07
	MM	0.31	0.25	0.23	0.19	0.16	0.13	0.11	0.1	0.09
	PWM	0.34	0.27	0.25	0.21	0.19	0.15	0.14	0.12	0.11
	GPWM	0.29	0.25	0.23	0.19	0.16	0.14	0.11	0.09	0.07

In Table 2, the reported RMSEs are in fact *relative* RMSEs, because they have been scaled by the true value of the quantile x_T. Shaded cells indicate the lowest RMSE obtained for each (k, n) combination.

The following observations can be made from Table 2: (1) The sample size does not seem to have a large effect on the relative performance of the different estimation methods with regard to RMSE of $\hat{x}_T$; (2) The GPWM method may be recommended for small k values, particularly for $k \leq -0.2$, which correspond to GPD distributions with extreme long tails; (3) The MM method may be recommended for $-0.3 < k \leq 0.1$, or even for k up to 0.2 if the sample size n is relatively large ($n \geq 50$); (4) The ML method may be recommended for larger values of k, such as for $k > 0.2$, which correspond to GPD distributions with extreme short tails and an upper bound. Note that the MM method could also have been recommended for higher values of k ($k > 0.2$, for example) on the basis of reported RMSEs of $\hat{x}_T$, especially when the sample size is small; however, for these higher k values the MM method suffers from a high rate of inconsistency with the data (Table 2). It is therefore more prudent to recommend the ML method for these higher *k* values. Note, in Table 2, that we have marked those cells that correspond to a high inconsistency rate, so that they can be easily identified.

Table 3 presents the RMSEs for $\hat{x}_T$, $T = 10$. Here, again, RMSEs have been scaled by the true value of x_T. The GPWM method is missing in Table 3 because Rasmussen's formulas for applying this method were provided only for $T \geq 50$.

Table 3 RMSEs of GPD quantile estimates corresponding to a return period $T = 10$. The notation is the same as in Table 2.

n	Method	*k*								
		-0.4	-0.3	-0.2	-0.1	0	0.1	0.2	0.3	0.4
15	ML	0.42	0.39	0.34	0.29	0.25	0.24	0.2	0.18	0.17
	MM	0.62	0.4	0.35	0.29	0.25	0.24	0.21	0.18	0.16
	PWM	0.38	0.35	0.33	0.29	0.25	0.24	0.21	0.18	0.17
25	ML	0.33	0.29	0.26	0.23	0.2	0.18	0.16	0.14	0.13
	MM	0.45	0.3	0.26	0.23	0.2	0.18	0.16	0.14	0.13
	PWM	0.3	0.27	0.25	0.23	0.2	0.19	0.16	0.14	0.13
50	ML	0.22	0.2	0.19	0.16	0.14	0.13	0.11	0.1	0.09
	MM	0.28	0.2	0.19	0.16	0.14	0.13	0.11	0.1	0.09
	PWM	0.21	0.19	0.19	0.16	0.14	0.13	0.11	0.1	0.09
100	ML	0.16	0.14	0.13	0.11	0.1	0.09	0.08	0.07	0.06
	MM	0.2	0.14	0.13	0.11	0.1	0.09	0.08	0.07	0.06
	PWM	0.16	0.14	0.13	0.11	0.1	0.09	0.08	0.07	0.06

The following observations can be made from Table 3: (1) There is very little difference between the three methods for k $\geq$ -0.2; (2) The PWM method has a slight advantage for k < -0.2 (distribution with extreme long tail); (3) for k > 0.2 the ML method may be recommended because it is the only method that does not suffer from a high rate of inconsistency with the data.

Conclusion and Recommendations

It has been shown that the MM, PWM and GPWM methods may produce estimates of the GPD upper bound that are inconsistent with the observed data. This issue was first pointed out by Dupuis (1996) but has been largely overlooked in the hydrological literature. We have shown that the "inconsistency rates" exhibited by any of the three methods just mentioned, can be as high as 30 %. Due to the seriousness of this problem, some of the main comparisons between the ML, MM, PWM and GPWM methods that have been previously reported in the hydrological literature, had to be re-examined. It was also noted that the ML method does not suffer from the problem of inconsistency with the simulated data that was displayed by the MM, PWM and GPWM methods.

For estimating GPD quantiles x_T with return period that is greater than the sample size (e.g., $T = 100$; $n = 15$, 25 or 50), the following recommendations may be made: (1) the GPWM method may be recommended when it is suspected that the population has an extreme long tail (e.g., $k \leq -0.2$); (2) The MM method may be recommended when there is evidence that the population has neither an extreme long tail nor an extreme short tail (e.g., $-0.3 < k \leq 0.1$); (3) The ML method may be recommended for larger values of k, such as for $k > 0.2$, which correspond to GPD distributions with extreme short tail.

For estimating a GPD quantile x_T with a return period that is smaller than the sample size it was shown that there is very little difference between the various methods. However, the ML method may be recommended because it is the only method that does not suffer from a high rate of inconsistency with the data.

References

Choulakian, V., and Stephens, M. A. (2001). "Goodness-of-fit for the Generalized Pareto distribution." *Technometrics,* 43, 478-484.

Dupuis, D. J. (1996). "Estimating the probability of obtaining nonfeasible parameter estimates of the generalized Pareto distribution." *Journal of Statistical Computation and Simulation,* 54, 197-209.

Hosking, J. R. M., and Wallis, J. R. (1987). "Parameter and quantile estimation for the generalized Pareto distribution." *Technometrics,* 29(3), 339-349.

Rasmussen, P. F. (2001). "Generalized probability weighted moments: Application to the generalized Pareto distribution." *Water Resources Research,* 37(6), 1745-1751.