Risk assessment using a generalized Pareto-based bivariate model

F. Ashkar¹ & N. El-Jabi²
¹Department of Mathematics and Statistics, Université de Moncton, Canada.
²Faculty of Engineering, Université de Moncton, Canada.

Abstract

In hydrology, one often has to study the joint probabilistic behavior of two or more variables characterizing a water system. A model constructed from generalized-Pareto-distributed (GPD) marginals is presented in the present paper to study the relation between two variables. The GPD model is the most popular one used in the 'Peaks-Over-Threshold' (POT) approach for studying hydrological extremes, and is also applicable in the 'Deficit-Below-Threshold' (DBT) approach for modeling low extremes. The bivariate model presented herein, allows for risk assessments and calculations associated with each hydrological variable separately, with one variable conditioned by the other, or with both variables taken together. It will be used to model the joint distribution of the duration (X₁) and the peak discharge above a threshold (X₂) of a real flood series derived from hydrometric data of the Little Southwest Miramichi River, in New Brunswick, Canada.

1 Introduction

In hydrology, one must often consider water systems where several random variables act individually or jointly on the system. For example, one must simultaneously take into account the intensity, duration and frequency of precipitation events when studying urban drainage systems. Likewise, the duration as well as the volume of low flows must be taken into account in managing water reservoirs, or in irrigation. It is important to be able to adequately describe the joint probability laws that govern the variables acting within the system.
Several hydrological publications have dealt with bivariate or multivariate statistical methods and their applications in describing relations between variables. Goel et al. [10], for example, used a bivariate distribution with exponential marginals to study rainfall intensity and duration, whereas Clarke [6] used a bivariate model to extend annual streamflow records from precipitation data. On the other hand, Ashkar et al. [1] described the joint probability law of two variables with marginal distributions having three parameters, where the joint probability density function (pdf) of the variables was obtained by transforming them into Gaussian variables via the Box-Cox transformation (Box and Cox [4]). This model was applied in a case study, where the variables $X_1$ and $X_2$ were annual maximum discharges, at two neighboring hydrometric stations, on two different streams.

In addition to their usefulness in studying high stream-flows, which has been demonstrated by Choulakian et al. [5] and Yue et al. [12], bivariate models have been used to study low stream-flows (Ashkar et al. [2]). Analysis of the joint distribution of various high and low flow characteristics, such as duration, volume or severity of high or low events, makes it possible to make good forecasts of these important hydrological variables.

In certain types of hydrological investigations, it may be appropriate to separate a variable being studied into two or more random components that are correlated or uncorrelated among each other, and to study these components separately, or jointly, using bivariate methods. Ashkar and Rousselle [3] performed such separation in a study dealing with peak flood discharge, flood duration and flood volume. These authors found it appropriate to divide the variable 'flood duration', $D$, into two components, $D_r$ and $D_f$, where $D_r$ is the rise duration, and $D_f$ is the fall duration. They found it reasonable to separate these two components, because the rising limb and the falling limb of a flood hydrograph are mainly affected by different factors. In fact, the dominant factor determining the shape of the rising limb of a flood is the character of the storm causing the rise, whereas the shape of the falling limb is mainly dependent on the basin characteristics during the time of flow recession, and largely independent of the storm that caused the flow rise. The variables $D_r$ and $D_f$ may therefore be considered as independent or almost independent of one another.

In some simple situations, it might be valid to assume independence between the random variables involved, in which case their joint distribution is easy to obtain. In more common situations, however, this independence assumption does not apply, which means that more complex joint distributions are needed to describe the phenomenon under investigation. It is important to choose the probability distribution that adequately models each variable separately, and also to take adequate account of the correlation that exists between the variables.

In the following sections we present a bivariate model based on the generalized Pareto distribution (GPD), as marginal for the two variables $X_1$ and $X_2$. It is to be noted that the GPD model is the most popular one used in the 'Peaks-Over-Threshold' (POT) approach for studying hydrological extremes (very high values of a hydrological variable). It is also a model applicable in the 'Deficit-Below-Threshold' (DBT) approach for modeling low extremes. These two
threshold-based methods, which are based on truncating an observed time series at a high or low level, are very useful mathematical tools for hydrological design.

2 Bivariate model with GPD marginals

A generalized Pareto random variable \( X \) may be defined as \( X = \alpha \left(1 - e^{-kY}\right)/k \), where \( Y \) is a random variable following a standard Exponential distribution. The pdf of \( X \) is given by

\[
f(x; \alpha, k) = \frac{1}{\alpha} \left(1 - k \frac{x}{\alpha}\right)^{1/k-1} \quad k \neq 0,
\]

\[
f(x; \alpha) = \frac{1}{\alpha} e^{-x/\alpha} \quad k = 0,
\]

whereas its cumulative distribution function (cdf) is given by

\[
F(x; \alpha, k) = 1 - \left(1 - k \frac{x}{\alpha}\right)^{1/k} \quad k \neq 0,
\]

\[
F(x; \alpha) = 1 - e^{-x/\alpha} \quad k = 0,
\]

where, for \( k < 0, 0 \leq x < \infty \), and for \( k > 0, 0 \leq x < \alpha/k \). The parameter \( \alpha \) is a scale parameter of \( X \), while \( k \) is a shape parameter. When \( k = 0 \), \( X \) reduces to an Exponential random variable with parameter \( \alpha \), whereas when \( k = 1 \), it reduces to a Uniform variable on the interval \((0, \alpha)\). The mean and variance of \( X \) are respectively given by

\[
\mu = E(X) = \frac{\alpha}{1+k}, \quad k > -1
\]

\[
\sigma^2 = Var(X) = \frac{\alpha^2}{(1+k)^2(1+2k)}, \quad k > -\frac{1}{2}
\]

2.1 Joint density

When \( X \) has pdf (1), we shall use the notation: \( X \sim GPD (\alpha, k) \). Let \( X_1 \sim GPD (\alpha_1, k_1) \) and \( X_2 \sim GPD (\alpha_2, k_2) \). Singh and Singh [11] presented the following joint density function of \( X_1 \) and \( X_2 \), based on work by Finch and Groblicki [9] and Cohen [7]:

\[
f(x_1, x_2) = f_1(x_1)f_2(x_2) \left[1 + c \rho \{F_1(x_1), F_2(x_2)\}\right]
\]
where $F_1(x_1)$ and $F_2(x_2)$ are the marginal cdfs of $X_1$ and $X_2$, $f_1(x_1)$ and $f_2(x_2)$ are the marginal densities, and $(u, v); u=F_1(x_1), v=F_2(x_2)$; is a function defined on the unit square $\{(u, v): 0 \leq u, v \leq 1\}$, whose calculation will be demonstrated in the next paragraphs. The constant ‘c’ in eqn (5) has to be selected so that the function $f(x_1, x_2)$ is positive.

To obtain the function $(u, v)$, a function $r(u, v)$ on the unit square $\{(u, v): 0 \leq u, v \leq 1\}$ has first be chosen such that

$$\int_0^1 \int_0^1 r(u, v) \, dudv = 1 \tag{6}$$

Many forms of the function $r(u, v)$ may be constructed, but one that we believe to be sufficiently flexible will be employed in the present study. In fact, Singh & Singh [11] investigated three different forms of the function $r(u, v)$, but all of these are special cases of the one chosen herein, which is:

$$r(u, v) = (n + 1)(m + 1) u^n v^m \tag{7}$$

where $m$ and $n$ are parameters to be chosen to fit the data under investigation. From this function, the association function $(u, v)$ between $u$ and $v$ is first calculated as follows (see, e.g., Singh & Singh [11]):

$$\rho(u,v) = r(u,v) - r_1(u) - r_2(v) + 1 \tag{8}$$

where

$$r_1(u) = \int_0^1 r(u,v) \, dv = \int_0^1 (n + 1)(m + 1) u^n v^m \, dv = (n + 1) u^n \tag{9}$$

$$r_2(v) = \int_0^1 r(u,v) \, du = \int_0^1 (n + 1)(m + 1) u^n v^m \, du = (m + 1)v^m \tag{10}$$

which gives,

$$\rho(u,v) = (n + 1)(m + 1) u^n v^m - (n + 1) u^n - (m + 1)v^m + 1$$

whence, the joint pdf of $(X_1, X_2)$ is given by
\[ f(x_1, x_2) = f_1(x_1)f_2(x_2)[1+c\rho(u,v)] \]
\[ = f_1(x_1)f_2(x_2)[1+c\left(\frac{n}{n+1}u^n - 1\right)\left(\frac{m}{m+1}v^m - 1\right)] \tag{12} \]

### 2.2 Joint cdf

The joint cdf of \((X_1, X_2)\) is given by

\[ F(x_1, x_2) = \int_{0}^{x_1} \int_{0}^{x_2} f(x_1, x_2) \, dx_1 \, dx_2 \]
\[ = \int_{0}^{x_1} \int_{0}^{x_2} f_1(x_1)f_2(x_2)[1+c\left(\frac{n}{n+1}u^n - 1\right)\left(\frac{m}{m+1}v^m - 1\right)] \, dx_1 \, dx_2 \]
\[ = \int_{0}^{x_1} \int_{0}^{x_2} [1+c\left(\frac{n}{n+1}u^n - 1\right)\left(\frac{m}{m+1}v^m - 1\right)] \, du \, dv \]
\[ = uv + c\int_{0}^{x_1} \left[\left(\frac{n}{n+1}u^n - 1\right)\left(\frac{m}{m+1}v^m - 1\right)\right] du \]
\[ = uv + c\left(\frac{n}{n+1}u^n - 1\right)\left(\frac{m}{m+1}v^m - 1\right) \]
\[ = F_1(x_1)F_2(x_2) + c\left[F_1(x_1)\left(\frac{n}{n+1}u^n - 1\right)\right] - F_1(x_1)\left[\left(\frac{m}{m+1}v^m - 1\right)\right] - F_2(x_2) \]  
\tag{13}

### 2.3 Conditional pdf and cdf

As for conditional pdf and cdf of \(X_1\), given \(X_2=x_2\), these are respectively determined as follows:

\[ f_{1\mid X_2}(x_1) = f_1(x_1)[1+c\rho(u,v)] \]
\[ F_1(x_1 \mid x_2) = \int_{0}^{x_1} f_{1\mid X_2}(x_1) \, dx_1 = \int_{0}^{x_1} f_{1\mid X_2}(x_1) + c f_{1\mid X_2}(x_1) \rho(u,v) \, dx_1 \]
\[ = f_1(x_1)dx_1 + c f_1(x_1)\rho(u,v)dx_1 \]
\[ F(x_i) + c \int_0^x f_i(x_i) [(n+1)u^n - 1] [(m+1)v^m - 1] dx_i \]

\[ F(x_i) + c [(m+1)v^m - 1] \int_0^x f_i(x_i) [(n+1)u^n - 1] dx_i \]

\[ F(x_i) + c [(m+1)v^m - 1] \int_0^x [(n+1)u^n - 1] du \]

\[ F(x_i) + c [(m+1)v^m - 1] \int_0^x [(n+1)u^n - 1] \left( u^{n+1} - u \right) du \]

\[ F(x_i) + c [(m+1) \left( \int f_2(x_2) v^m - 1 \right) \left( \int f_1(x_1) u^n - 1 \right) - F_1(x_i) ] \]

Equation (15)

And, as mentioned earlier, the constant ‘c’ should be chosen so that \( f(x_1, x_2) \) is positive. This means that for the association function \( \rho \) of eqn (11) we need to have

\[ -\frac{1}{\max(1, mn)} \leq c \leq \frac{1}{\max(m, n)} \]

Equation (16)

3 Case study

The bivariate distribution with GPD marginals presented above will be used to model the joint distribution of the duration \( (X_1) \) and the peak discharge above a threshold \( (X_2) \) of a real flood series. The hydrometric data used are those of the Little Southwest Miramichi River at Lyttleton (Station BP001), in New Brunswick, Canada. The basin area is of 1340 km². A threshold level of 175 m³/s was chosen, which represents the 98th percentile of daily river-flows, classified in ascending order. The period of record is of 45 years (1952-1996), during which \( N = 81 \) flood events exceeding the threshold were observed.

First, the parameters \( (\alpha_1, k_1, \alpha_2, \text{and } k_2) \) of the marginal pdf’s \( f_1(x_1) \) and \( f_2(x_2) \) (eqn (1)) were estimated using the method of moments (Dupuis, [8]). These parameter estimates are presented in Table 1, along with some descriptive statistics for the \( x_1 \) and \( x_2 \) data series.

Following estimation of the parameters of the marginal pdf’s, the functions \( f(x_1, x_2), F(x_1, x_2), f(x_1 | x_2) \) and \( F(x_1 | x_2) \) had to be estimated. To estimate the parameters ‘m’ ‘n’ and ‘c’, a 3x3 contingency table was constructed from the \( N = 81 \) observed couples \( (x_1, x_2) \). Table 2 presents the cell counts forming this contingency table. The construction of this contingency table was done so that the row and column total counts (marginal counts) are almost uniformly distributed. For simplification purposes, we took \( m = n \) in eqns (12) – (15), and proceeded to find the triplet \( (n, m, c; n = m) \) that minimizes the chi-squared distance between observed cell counts \( f_{\text{obs}} \) and estimated cell counts \( f_{\text{est}} \) under the hypothesized model. This chi-squared distance is given by:
Table 1: Parameter estimates for the model and some descriptive statistics for the $x_1$ and $x_2$ data series.

<table>
<thead>
<tr>
<th>Descriptive statistics</th>
<th>$x_1$ (days)</th>
<th>$x_2$ (m$^3$/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.1</td>
<td>80.0</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>2.6</td>
<td>115.2</td>
</tr>
<tr>
<td>Median</td>
<td>2.0</td>
<td>41.0</td>
</tr>
<tr>
<td>Minimum</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Maximum</td>
<td>13</td>
<td>686</td>
</tr>
<tr>
<td>Correlation ($x_1, x_2$)</td>
<td>0.52</td>
<td></td>
</tr>
</tbody>
</table>

Estimated parameters

<table>
<thead>
<tr>
<th>$k$</th>
<th>0.251</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$</td>
<td>-0.259</td>
</tr>
<tr>
<td>$n$</td>
<td>0.726</td>
</tr>
<tr>
<td>$M$</td>
<td>1.378</td>
</tr>
</tbody>
</table>

Table 2: Contingency table for the $N = 81$ observed couples ($x_1$, $x_2$), $x_1$ in days, $x_2$ in m$^3$/s.

<table>
<thead>
<tr>
<th>0 &lt; $x_1$ ≤ 1</th>
<th>1 &lt; $x_1$ ≤ 3</th>
<th>3 &lt; $x_1$ &lt; ∞</th>
<th>Row Marginal</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 &lt; $x_2$ ≤ 28</td>
<td>19</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>28 &lt; $x_2$ ≤ 59</td>
<td>7</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>59 &lt; $x_2$ &lt; ∞</td>
<td>3</td>
<td>6</td>
<td>18</td>
</tr>
</tbody>
</table>

| Column Marginal Total | 29 | 25 | 27 | 81 |

$$\chi^2 = \sum_{i=1}^{9} \left( \frac{f_{obs} - f_{th}}{f_{th}} \right)^2$$  (17)

The values of $n$, $m$ and $c$ that minimize this chi-squared distance are presented in Table 1. All the parameters of the model having now been estimated, it is now possible to test the fit of the model to the data. One way to examine this fit is to plot the empirical and estimated conditional cdfs $F(x_1 | x_2)$ of flood duration conditioned by peak discharge above the threshold.

Figures 1 to 3 show the empirical and estimated (“theoretical”) cdfs $F(x_1 | x_2)$ of flood durations conditioned by peak flood exceedances over the threshold. Figure 1 is based on the subset of observed couples ($x_1$, $x_2$) characterized by a “low $x_2$ value” ($0 < x_2 ≤ 28$ m$^3$/s, representing the first row of
Figure 1: Empirical and theoretical cdfs $F(x_1 \mid x_2 = 12 \text{ m}^3/\text{s})$ of flood durations conditioned by peak exceedances over the threshold.

Figure 2: Empirical and theoretical cdfs $F(x_1 \mid x_2 = 41 \text{ m}^3/\text{s})$ of flood durations conditioned by peak exceedances over the threshold.
Figure 3: Empirical and theoretical cdf's $F(x_1 \mid x_2 = 120 \text{ m}^3/\text{s})$ of flood durations conditioned by peak exceedances over the threshold.

Table 2, with a median equal to 12 m$^3$/s). Similarly, Figures 2 and 3 are based, respectively, on observed couples $(x_1, x_2)$ characterized by a “medium $x_2$ value” (28 < $x_2$ ≤ 59 m$/^3$/s; median = 41 m$^3$/s), and a “high $x_2$ value” (59 < $x_2$ < $\infty$ m$^3$/s; median = 120 m$^3$/s). The theoretical curve in Figure 1 is obtained by plugging the value $x_2 = 12$ m$^3$/s into eqn (15), whereas the corresponding values used in Figures 2 and 3 are respectively $x_2 = 41$ m$^3$/s and $x_2 = 120$ m$^3$/s. The empirical curves in Figures 1 to 3 are obtained using a plotting formula.

From Figures 1 to 3 it is seen that the model provides a reasonably good fit to the Little Southwest Miramichi River data. Note, however, that the fit in Figure 1 could most probably be significantly improved if the flood duration data were available to the nearest hour instead of to the nearest day.

Acknowledgments

The financial support of the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged. Miss Lampouguin Bayentin assisted in calculations related to fitting of probability distributions.

References


