The Weibull distribution applied to regional low flow frequency analysis

P. J. PILON
Environment Canada, Inland Waters Directorate, Water Resources Branch, Ottawa, Ontario, Canada K1A 0H3

Abstract The inability to estimate accurately low flows of specific duration and probability for ungauged basins has long plagued the practitioner. The index-flood method is one tool which may assist in this regard, through its adaptation to low flows. This paper outlines the extension of the index-flood method to low flow analysis when the regional distribution is assumed to be the three-parameter Weibull. The derivation of a homogeneity test is also given. The approach is demonstrated by its application to an area in the southern portion of Ontario, a province of Canada.

Application de la distribution de Weibull à l'analyse régionale des débits d'étiage

Résumé Les spécialistes se préoccupent depuis longtemps d'être en mesure de produire des estimations précises des débits d'étiage de probabilité et durée spécifiques pour les bassins non jaugés. La méthode des crues repères (index-flood method) constitue un outil qui peut s'avérer utile à cet égard lorsqu'elle est adaptée aux débits d'étiage. Le présent article donne les grandes lignes de l'extension de la méthode des crues repères à l'analyse des débits d'étiage en supposant que la distribution régionale est une distribution de Weibull à trois paramètres. La dérivation du test d'homogénéité est également fournie. L'approche est illustrée par son application dans une région au sud de l'Ontario, province canadienne.

INTRODUCTION

Knowledge of a stream's low flow characteristics is important for several engineering applications, such as environmental impact assessment, waste-water effluent dilution, and water supply for potable and irrigation purposes. The relation of the magnitude of the low flow with its frequency of occurrence provides some insight as to the general availability of water. This relation is usually obtained from the application of a theoretical probability distribution to hydrometric data.

Unfortunately, hydrometric gauging records are not always available at the site of interest. Hence, methods have been developed to transfer
widely-used streamflow indexes from gauged to ungauged watersheds. Extensive work in this area has been done on the frequency of floods. Two methods have emerged:

(a) the direct regression approach based on ordinary, weighted, or generalized least squares (Benson, 1962; Thomas & Benson, 1970; Stedinger & Tasker, 1985);

(b) the index-flood approach (Dalrymple, 1960; Harvey et al., 1985).

Both methods require the delineation of homogeneous regions for which their models are applicable. For the direct regression approach, the regions are commonly obtained by mapping the models' residuals. Patterns are visually assessed, with homogeneous regions being geographically assigned to minimize bias. New models are then formulated for these regions, and the residual patterns are again assessed. This process continues until the residuals indicate no bias in the model.

The index-flood procedure is composed of two major parts. The first is the dimensionless frequency curve that represents the ratio of flood discharges at selected recurrence intervals to an index flood. The curve is applicable within a homogeneous region defined such that the slopes of the single site curves do not vary more than that due to random sampling for the assumed parent distribution. A homogeneity test is used to study the slopes of the individual frequency curves. The second major part of this approach is the development of a model to estimate the index flood. It usually is a least squares model and is obtained using the direct regression procedure. The combination of the index-flood model with the dimensionless frequency curve yields a frequency curve at any location.

Wiltshire (1986) and Bhaskar & O'Connor (1989) describe an alternative approach to that outlined above. Regions are defined based on a cluster analysis of two statistics of gauged data. The goal is to obtain regions with basins that have comparable flood frequency characteristics and that may have similar geomorphology. This definition of homogeneity is not geographically based, but is based on flood characteristics. This represents a departure from the classical two-dimensional or mapping view of homogeneity to the multivariate level. This approach could potentially assist regionalization studies involving the two methods.

Both methods are popular in flood studies. In low flow studies, the direct regression method is usually applied to low flow indexes such as the 10-year nonexceedance, 7-day mean low flow. However, it has not had great success due primarily to the inaccuracies of the derived least squares model.

The purpose of this paper is to describe a procedure analogous to the index-flood approach, but directed towards low flows. The index-low flow procedure is based on one of the most commonly used theoretical distributions in the single site, frequency analysis of low flows. The distribution is the three parameter Weibull (W3), which is also commonly referred to in hydrology as the Gumbel III distribution. A homogeneity test for the slopes of the frequency curves is derived. The method is applied to an area of southern Ontario for demonstration purposes.
CONSIDERATION OF THE DISTRIBUTION

Several studies have investigated various methods for estimating the low flow characteristics of streams (Matalas, 1963; Condie & Nix, 1975; Loganathan et al., 1985; and Tasker, 1987). These studies were concerned with comparing or describing methods to relate the magnitude of low flows with their frequency of occurrence.

The studies of Matalas (1963) and Condie & Nix (1975) were limited to distribution functions that are theoretically lower bounded. They felt that a distribution's lower boundary should be located between zero and the smallest observed low flow of the sample. This was similar in philosophy to the work of Weibull on the breaking strengths of materials, as described by Gumbel (1958). Condie & Nix (1975) reported that the W3 distribution yielded lower boundary parameters between zero and the smallest observed flow for 33 out of 38 Canadian rivers. Matalas (1963) reported similar results for 31 of 34 rivers in the United States of America. Tasker (1987), using bootstrapping, found that the log Pearson type III and W3 distributions tended to be more accurate in estimating the low flow frequency regime than did the Box-Cox transformation and log-Boughton methods.

Loganathan et al. (1985) questioned the use of both the W3 and log Pearson type III distribution in low flow frequency analysis, as both "can have positive lower bounds and thus their use in fitting low flow probabilities may not be justifiable." Gumbel (1958) felt that in practical application that the lower boundary or location parameter was the most important of the three parameters. He cautioned, however, that estimation procedures sometimes yield boundary estimates larger than the smallest observation. Such an estimate of the lower boundary cannot be used. This led to the formulation of alternative parameter estimation methods such as the smallest observed drought, the results of which are more physically realistic, but not necessarily more accurate.

The potential problems associated with the use of a lower-bounded distribution such as the W3 for the assessment of the probabilities of nonexceedance of low flows did not deter either theoretician or practitioner. The result has been a widespread adoption of the W3 as the parent distribution of low flows in several studies. Thus, the W3 distribution is a strong candidate for a regional index procedure.

THE THREE PARAMETER WEIBULL DISTRIBUTION

A detailed description of the distribution and several commonly used methods for the estimation of its parameters is not presented herein as it is available in several publications (Gumbel, 1958; Matalas, 1963; Condie & Nix, 1975; Deininger & Westfield, 1969; Kite, 1976). Only the probability density function will be provided, as it is used in the derivations which follow and in the construction of the homogeneity test. The probability density function of the W3 distribution is:
\[
\phi(x) = \frac{a}{u - e} \left( \frac{x - e}{u - e} \right)^{a - 1} \exp \left[ -\left( \frac{x - e}{u - e} \right)^a \right]
\]  
(1)

where \( e \) is the lower boundary parameter, \( u \) is the characteristic drought, and \( a \) is the shape parameter. The density function is integrable and gives:

\[
F(x) = 1 - \exp \left[ -\left( \frac{x - e}{u - e} \right)^a \right]
\]  
(2)

where \( F(x) \) is the probability of nonexceedance of \( x \) and is the inverse of the return period of nonexceedance, \( T \). \( x \) can be obtained by rearranging equation (2) to give:

\[
x = e + (u - e)(-\ln[1 - F(x)])^{1/a}
\]  
(3)

**REGIONALIZATION**

For every station, a low flow frequency analysis is performed based on the W3 distribution. Flows corresponding to specific return periods of nonexceedance can be made dimensionless by dividing by some chosen index low flow. The 2-year nonexceedance low flow is used herein.

Within a homogeneous region, the dimensionless frequency curve at any station is considered a random sample. The best representation of the regional characteristics is obtained by averaging the dimensionless curves for all stations in the region. The resulting average dimensionless curve is the regional dimensionless frequency curve and is considered applicable throughout the region, providing the conditions of homogeneity are met. If the 2-year low flow at an ungauged site can be estimated, the entire low flow frequency relationship can be developed by multiplying by the appropriate ratios of the dimensionless curve.

From equation (3), any three low flows and their return periods of nonexceedance give three simultaneous transcendental equations which, when solved, yield parameters \( a \), \( e \) and \( u \). If the 2, 12.488, and 100-year low flows are selected, then the solution is simplified to the evaluation of three functions.

The procedure then is to find the median values of the dimensionless 12.488 and 100-year low flows in the region. The value 12.488 is the mid-point between the 2 and 100-year return period in Gumbel reduced variate terms. Note that the median dimensionless value of all the 2-year low flows is unity. The medians are then substituted into the following expressions to obtain the parameters of the dimensionless curve:

\[
e = (Q_{100} - Q_{12.488}^2)/(1 + Q_{100} - 2Q_{12.488})
\]  
(4)

\[
a = 4.23464/\ln[(1 - e)/(Q_{100} - e)]
\]  
(5)
\[ u = \left[ (1 - e)/0.69315^{1/a} \right] + e \]  

(6)

where \( Q_{100} \) and \( Q_{12.488} \) represent the median values of the 100 and 12.488-year indexed low flows. The regional dimensionless low flows follow from these parameters and equation (3).

**HOMOGENEITY TEST**

The test used herein is similar to that described by Dalrymple (1960). For each station, the 10-year nonexceedance flow is tested by multiplying its estimated 2-year nonexceedance flow by the regional \( Q_{100} \) index ratio. Using the station analysis, the return period, \( T \), of the estimated 10-year nonexceedance flow, \( x_{10} \), can be found from:

\[ T = 1/(1 - \exp[-\exp(a \ln((x_{10} - e)/(u - e)))]) \]  

(7)

The value of \( T \) will almost certainly never be 10 years. However, \( T \) should fall within two standard errors of the return period for a given sample size \( n \) to be expected in random sampling from the W3 distribution. If we apply a change of variate and let:

\[ y = [(x - e)/(u - e)]^a \]  

(8)

then

\[ dy = a \left[ \frac{x - e}{u - e} \right]^{a-1} \frac{1}{u - e} \, dx \]  

(9)

and

\[ \phi(y) = \phi(x) \frac{dx}{dy} = \exp(-y) \]

The retention of a scaling parameter, \( c \), for the interim will assist in the derivation. Then, if \( y/c \) is a gamma variate, parameter \( b \), with \( b \) set to one:

\[ \phi(y) = c^{-1} \exp(-y/c) \]  

(10)

where \( c \) is the mean.

The logarithmic likelihood function, \( \ln L \), of equation (10) is:

\[ \ln L = -(1/c) \Sigma y - n \ln c \]  

(11)

where \( n \) represents the sample size over which the summation, \( \Sigma \), is taken. Following the work of Fisher, the variance of \( c \) can be estimated as:
The probability of a value \( y \) being less than \( y_c \) is:

\[
F(y_c) = \int_0^{y_c} e^{-y/c} \frac{dy}{c} = 1 - \exp(-y_c/c)
\]  

(13)

Rearranging this equation gives:

\[
y_c = -c \ln(1 - 1/T)
\]  

(14)

where \( y_c \) is the \( T \)-year nonexceedance flow.

The variance of \( y_c \) can be given by the expression:

\[
\text{var}(y_c) = (\frac{\partial y_c}{\partial c})^2 \text{var}(c)
\]  

(15)

or

\[
\text{var}(y_c) = c^2 \frac{[-\ln(1 - 1/T)]^2}{n}
\]

The square root of equation (15) yields the standard error of estimate of the \( T \)-year nonexceedance flow. So, for the 10-year nonexceedance flow and letting \( c \) be one,

\[
\sigma_{y_c} = 0.10536/\sqrt{n}
\]  

(16)

From this relationship the approximate upper and lower 95% confidence limits can be computed for any sample size \( n \). Table 1 lists the limits for various sample sizes. A funnel-shaped control chart could be constructed with the approximate upper and lower 95% confidence limits from Table 1. The sample size, \( n \), would be plotted as the abscissa and \( T_L \) and \( T_U \) as the

<table>
<thead>
<tr>
<th>( n ) (years)</th>
<th>( Q_{10} )</th>
<th>( \sigma = \frac{0.21}{n^{0.21}} )</th>
<th>( Q_{10} + 2\sigma )</th>
<th>( Q_{10} - 2\sigma )</th>
<th>( T_L )</th>
<th>( T_U )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.10536</td>
<td>0.067</td>
<td>0.172</td>
<td>6.33</td>
<td>0.039</td>
<td>26.3</td>
</tr>
<tr>
<td>15</td>
<td>0.10536</td>
<td>0.054</td>
<td>0.160</td>
<td>6.77</td>
<td>0.051</td>
<td>20.1</td>
</tr>
<tr>
<td>20</td>
<td>0.10536</td>
<td>0.047</td>
<td>0.152</td>
<td>7.07</td>
<td>0.058</td>
<td>17.7</td>
</tr>
<tr>
<td>25</td>
<td>0.10536</td>
<td>0.042</td>
<td>0.148</td>
<td>7.29</td>
<td>0.063</td>
<td>16.3</td>
</tr>
<tr>
<td>30</td>
<td>0.10536</td>
<td>0.039</td>
<td>0.144</td>
<td>7.46</td>
<td>0.067</td>
<td>15.5</td>
</tr>
<tr>
<td>40</td>
<td>0.10536</td>
<td>0.033</td>
<td>0.139</td>
<td>7.77</td>
<td>0.072</td>
<td>14.4</td>
</tr>
<tr>
<td>50</td>
<td>0.10536</td>
<td>0.030</td>
<td>0.135</td>
<td>7.91</td>
<td>0.076</td>
<td>13.7</td>
</tr>
<tr>
<td>60</td>
<td>0.10536</td>
<td>0.027</td>
<td>0.133</td>
<td>8.05</td>
<td>0.078</td>
<td>13.3</td>
</tr>
<tr>
<td>80</td>
<td>0.10536</td>
<td>0.024</td>
<td>0.129</td>
<td>8.27</td>
<td>0.083</td>
<td>12.7</td>
</tr>
<tr>
<td>100</td>
<td>0.10536</td>
<td>0.021</td>
<td>0.126</td>
<td>8.42</td>
<td>0.084</td>
<td>12.4</td>
</tr>
</tbody>
</table>
ordinates, where $T_L$ and $T_U$ are the return periods associated with the lower and upper confidence limits, respectively.

**AN EXAMPLE**

This example is intended only for illustration purposes and does not constitute a regional low flow frequency study. Sangal & Kallio (1977) performed a comprehensive analysis of flooding in southern Ontario and subdivided the area into nine regions, as shown in Fig. 1. Region 3, the Saugeen-Nottawasaga region, is used in this study. An extension of the regional model to Region 5, Lake St Clair, demonstrates the sensitivity of the homogeneity test.

Table 2 lists all natural basins having more than 10 years of streamflow data and having an area greater than 200 km$^2$ in Regions 3 and 5. The areas of the drainage basins are listed in Table 2. Table 3 gives the dimensionless 7-day low flow ratios for the basins. The ratios are obtained from single station frequency analysis using the W3 distribution, where the parameters are estimated by the maximum likelihood approach. Station 02GG004 is the only site that requires an alternative estimation procedure — the smallest observed drought.

The median values of the ratios for Region 3 can be substituted into equations (4), (5) and (6). This gives the parameters $a$, $e$ and $u$ of the dimensionless regional 7-day low flow frequency curve to be 2.22943, 0.45588, and 1.09722, respectively. Substituting these parameter values into equation (3) yields the dimensionless ratio for any probability of nonexceedance. For example, the 10-year nonexceedance ratio is 0.6896 and is used in conjunction with the index low flow to estimate the regional 10-year low flow.

**Table 2** Drainage basins, areas, and record lengths used in the study

<table>
<thead>
<tr>
<th>W.S.C.* no.</th>
<th>Station name</th>
<th>Record length</th>
<th>Area, (km$^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>02ED003</td>
<td>Nottawasaga River near Baxter</td>
<td>37</td>
<td>1180</td>
</tr>
<tr>
<td>02FA001</td>
<td>Sauble River at Sauble Falls</td>
<td>29</td>
<td>927</td>
</tr>
<tr>
<td>02FB010</td>
<td>Bighead River near Meaford</td>
<td>29</td>
<td>293</td>
</tr>
<tr>
<td>02FC001</td>
<td>Saugeen River near Port Elgin</td>
<td>72</td>
<td>3960</td>
</tr>
<tr>
<td>02FC002</td>
<td>Saugeen River near Walkerton</td>
<td>72</td>
<td>2150</td>
</tr>
<tr>
<td>02FC015</td>
<td>Teeswater River near Paisley</td>
<td>14</td>
<td>663</td>
</tr>
</tbody>
</table>

REGION 5

<table>
<thead>
<tr>
<th>W.S.C.* no.</th>
<th>Station name</th>
<th>Record length</th>
<th>Area, (km$^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>02FF002</td>
<td>Ausable River near Springbank</td>
<td>39</td>
<td>865</td>
</tr>
<tr>
<td>02FF007</td>
<td>Bayfield River near Varna</td>
<td>19</td>
<td>466</td>
</tr>
<tr>
<td>02GG002</td>
<td>Sydenham River near Alvinston</td>
<td>38</td>
<td>730</td>
</tr>
<tr>
<td>02GG004</td>
<td>Bear Creek above Wilkesport</td>
<td>18</td>
<td>609</td>
</tr>
<tr>
<td>02GG007</td>
<td>Sydenham River near Dresden</td>
<td>16</td>
<td>1240</td>
</tr>
</tbody>
</table>

*W.S.C. implies the Water Survey of Canada.
Fig. 1 Flood frequency regions in southern Ontario (after Sangal & Kallio, 1977).
The Weibull distribution and low flow frequency analysis

Table 3 Dimensionless low flow ratios

<table>
<thead>
<tr>
<th>W.S.C.* no.</th>
<th>Return period (years):</th>
<th>2</th>
<th>12.488</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>REGION 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02ED003</td>
<td>1</td>
<td>0.67171</td>
<td></td>
<td>0.50693</td>
</tr>
<tr>
<td>02FA001</td>
<td>1</td>
<td>0.50467</td>
<td></td>
<td>0.40822</td>
</tr>
<tr>
<td>02FB010</td>
<td>1</td>
<td>0.65317</td>
<td></td>
<td>0.57341</td>
</tr>
<tr>
<td>02FC001</td>
<td>1</td>
<td>0.71911</td>
<td></td>
<td>0.60418</td>
</tr>
<tr>
<td>02FC002</td>
<td>1</td>
<td>0.66203</td>
<td></td>
<td>0.52433</td>
</tr>
<tr>
<td>02FC015</td>
<td>1</td>
<td>0.67081</td>
<td></td>
<td>0.55036</td>
</tr>
<tr>
<td>REGION 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02FF002</td>
<td>1</td>
<td>0.46819</td>
<td></td>
<td>0.30869</td>
</tr>
<tr>
<td>02FF007</td>
<td>1</td>
<td>0.49602</td>
<td></td>
<td>0.27880</td>
</tr>
<tr>
<td>02GG002</td>
<td>1</td>
<td>0.59423</td>
<td></td>
<td>0.50099</td>
</tr>
<tr>
<td>02GG004</td>
<td>1</td>
<td>0.03959</td>
<td></td>
<td>-0.06812</td>
</tr>
<tr>
<td>02GG007</td>
<td>1</td>
<td>0.56719</td>
<td></td>
<td>0.38263</td>
</tr>
</tbody>
</table>

* W.S.C. implies the Water Survey of Canada.

HOMOGENEITY TEST

Having derived the parameters of the regional dimensionless curve, the individual stations can now be tested for homogeneity using either equation (16) or Table 1. Table 4 gives the results of the derived regional model for Regions 3 and 5. The model based on the data of Region 3 yields nonhomogeneity at two of the six sites. Application of the model to Region 5 indicates considerable bias and nonhomogeneity. Table 4 also lists the percentage error of the estimate of the 10-year nonexceedance flow.

Table 4 Application of homogeneity test

<table>
<thead>
<tr>
<th>W.S.C. no.</th>
<th>10-year at site discharge (m³ s⁻¹)</th>
<th>10-year regional discharge (m³ s⁻¹)</th>
<th>% error</th>
<th>T of regional discharge (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>02ED003</td>
<td>1.549</td>
<td>1.531</td>
<td>-1.2</td>
<td>10.70</td>
</tr>
<tr>
<td>02FA001</td>
<td>0.643</td>
<td>0.839</td>
<td>30.5</td>
<td>4.11*</td>
</tr>
<tr>
<td>02FB010</td>
<td>0.351</td>
<td>0.361</td>
<td>2.8</td>
<td>8.22</td>
</tr>
<tr>
<td>02FC001</td>
<td>8.195</td>
<td>7.648</td>
<td>6.7</td>
<td>18.14*</td>
</tr>
<tr>
<td>02FC002</td>
<td>4.363</td>
<td>4.384</td>
<td>0.5</td>
<td>9.72</td>
</tr>
<tr>
<td>02FC015</td>
<td>0.931</td>
<td>0.926</td>
<td>-0.5</td>
<td>10.34</td>
</tr>
<tr>
<td>REGION 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02FF002</td>
<td>0.144</td>
<td>0.199</td>
<td>37.9</td>
<td>4.13*</td>
</tr>
<tr>
<td>02FF007</td>
<td>0.061</td>
<td>0.079</td>
<td>28.9</td>
<td>4.70*</td>
</tr>
<tr>
<td>02GG002</td>
<td>0.430</td>
<td>0.482</td>
<td>12.1</td>
<td>5.69*</td>
</tr>
<tr>
<td>02GG004</td>
<td>0.0036</td>
<td>0.015</td>
<td>853</td>
<td>2.56*</td>
</tr>
<tr>
<td>02GG007</td>
<td>0.527</td>
<td>0.608</td>
<td>15.3</td>
<td>5.87*</td>
</tr>
</tbody>
</table>

*implies outside the 95% confidence limits.
Apparently, an error of as low as 7% (02FC001) can result in a site being classified nonhomogeneous, though such an error may be practically insignificant.

DISCUSSION

The application of the index-low flow procedure illustrates some of its drawbacks. Two basins within the region demonstrate nonhomogeneity and there is little possibility of subdividing the region further. The index approach has also fallen under some criticism as the 7-year event is always a constant multiple of the index low flow, regardless of basin characteristics. The observed nonhomogeneity and the criticism are due to the geographic basis for the demarcation of homogeneous regions. This premise for regionalization does not reflect the geomorphic diversity which can exist among basins of a geographic area. Regions defined by a basin's response under drought conditions, such as was done by Wiltshire (1986) and Bhaskar & O'Connor (1989), may be more physically meaningful and may increase the applicability and accuracy of the index-low flow procedure.

The success of regionalization depends on the ability to define homogeneous areas and on the accuracy and reliability of the regional dimensionless index curve. As the parent distribution is assumed a priori to be the W3, the density function is lower bounded at parameter $e$. The lower boundary at a particular site is $e$ times its 2-year nonexceedance flow. The probability of obtaining low flows less than the boundary is of course zero. Comparison of the extreme lower tail with single site data could show that the regional model yields physically impossible results at times. Under these circumstances, the ability of the selected distribution to adequately represent the distribution of low flow may be questioned.

When applying the derived index-low flow method, a primary assumption is that the parent distribution is W3. This premise should always be investigated, as it is entirely plausible that low flows of a region are not drawn from this parent. L-moment ratio diagrams could be used to study the appropriateness of the choice of the parent distribution (Wallis, 1988; Hosking, 1989). When the parent is W3, the derived technique may provide useful results in many practical situations.

REFERENCES


Harvey, K. D., Condie, R., & Pilon, P. J. (1985) Regional flood frequency analysis with the three-parameter lognormal distribution. *Proc. CSCE 7th Canadian Hydrotechnical Conference* (Saskatoon, Saskatchewan, 27 May-1 June).


