Wind and Structures, *Vol. 19, No. 4 (2014) 371-387*
DOI: http://dx.doi.org/10.12989/was.2014.19.4.371

Plotting positions and approximating first two moments of order statistics for Gumbel distribution: estimating quantiles of wind speed

H.P. Hong* and S.H. Li[a]

Department of Civil and Environmental Engineering, University of Western Ontario, Canada N6A 5B9

(Received October 22, 2013, Revised June 13, 2014, Accepted July 14, 2014)

Abstract. Probability plotting positions are popular and used as the basis for distribution fitting and for inspecting the quality of the fit because of its simplicity. The plotting positions that lead to excellent approximation to the mean of the order statistics should be used if the objective of the fitting is to estimate quantiles. Since the mean depends on the sample size and is not amenable for simple to use closed form solution, many plotting positions have been presented in the literature, including a new plotting position that is derived based on the weighted least-squares method. In this study, the accuracy of using the new plotting position to fit the Gumbel distribution for estimating quantiles is assessed. Also, plotting positions derived by fitting the mean of the order statistics for all ranks is proposed, and an approximation to the covariance of the order statistics for the Gumbel (and Weibull) variate is given. Relative bias and root-mean-square-error of the estimated quantiles by using the proposed plotting position are shown. The use of the proposed plotting position to estimate the quantiles of annual maximum wind speed is illustrated.

Keywords: Gumbel distribution; Weibull distribution; least-squares methods; plotting positions

1. Introduction

There are several popular distribution fitting methods, including the method of moments, the maximum likelihood method and the ordinary least-squares (OL) method (Jordaan 2005). The OL method that is a particular case of the generalized least-squares (GL) method is attractive partly due to its graphical appear and its ease to use. Lloyd (1952) showed that the best linear unbiased estimators (BLUE) for distribution fitting can be obtained using the generalized least-squares (GL) method using the mean and covariance matrix of the order statistics (David and Nagaraja 2003). This approach for the Gumbel distribution is presented by Lieblein (1953, 1974); the BLUE coefficients are provided but only for sample size less than or equal to 16 because of the difficulty in evaluating the moments of the order statistics. The relation between the Weibull and Gumbel distributions is exploited by Lieblein and Zelen (1956) to estimate the distribution parameters for the Weibull model. To simplify the analysis, Harris (1996) proposed to retain the variance and neglecting the covariance of the order statistics (i.e., to use the weighted least-squares (WL)

*Corresponding author, Professor, E-mail: hongh@eng.uwo.ca
[a] Ph.D. Student, E-mail: sli472@uwo.ca

http://www.techno-press.org/?journal=was&subpage=8 ISSN: 1226-6116 (Print), 1598-6225 (Online)

method) for distribution fitting. If the variation of the order statistics is neglected, the WL method becomes the OL method (David and Nagaraja 2003). In this case, only the mean of the order statistics is needed. Since there is no simple to use equation to evaluate the mean of the order statistics in many cases, further simplification is achieved by adopting plotting positions that approximate this mean. This indicates that the plotting positions that lead to excellent approximation to the mean of the order statistics should be considered in the OL method.

A comparison of the performance of several distribution fitting methods for the Gumbel distribution (Hong *et al*. 2013, Hong 2013) indicates that the GL method is preferred in terms of minimum bias and minimum root-mean-square-errors (RMSE), and that the differences between the estimated quantiles by using the GL, WL and OL methods decrease as the sample size increases.

An earlier review of various plotting positions and their adequacy is given by Cunnane (1978), who argued that "the plotting position should be such that an estimated quantile should be free from bias and should have minimum variance among graphical estimates"; the review given by Harter (1984) emphasized that the optimum plotting position depends on the objective of the analysis. Cunnane (1978) indicated that Gringorten (1963) formula is preferred for the Gumbel distribution. Since this formula is derived by assuming that the probability distribution is symmetric and by matching the mean of the largest order statistics, it could be deficient for the Gumbel distribution. By relaxing the symmetry assumption and by matching the means of the smallest and the largest order statistics, Cook and Harris (2013) showed the asymptotic behaviour of the plotting position. Using this as a guide, they proposed simple to use plotting positions for the Gumbel, Weibull and exponential distributions based on the weighted least-squares fits. They further suggested that the plotting position can be more appropriately derived by considering all ranks of the order statistics. In fact, such an approach was taken in Hong (2013) to suggest a new plotting position based on the estimated mean of the order statistics for the sample size up to 300. However, this suggested plotting position is less convenience to use than the one given by Cook and Harris (2013).

In this study, the accuracy of using the plotting positions given by Cook and Harris (2013) to fit the Gumbel distribution for estimating quantile is reported. Also, new plotting positions derived by fitting the mean of the order statistics for all ranks are given, and approximations to the variance and covariance of the order statistics for the Gumbel variate and Weibull variate are presented. The proposed plotting position is motivated by the simplicity of the one recommended in Cook and Harris. Relative bias and RMSE of the estimated quantiles by using the proposed plotting position as well as by using the approximate variance and covariance are shown. The use of the proposed plotting position to estimate the quantiles of annual maximum wind speed is illustrated.

2. Plotting position and approximation to the moments of order statistics for Gumbel distribution

2.1 Plotting positions and approximation to mean of order statistics

The Gumbel distribution function for the random variable X, $F_G(x)$, is given by (Castillo 1988, Jordaan 2005)

$$F_G(x) = \exp\left(-\exp\left(-\frac{x-u}{a}\right)\right) \tag{1}$$

where u and a are the location and scale parameters. Define the reduced variate, Y

$$Y = (X-u)/a \tag{2}$$

The probability distribution function of Y, $F_Y(y)=\exp(-\exp(-y))$, and its inverse $F_Y^{-1}(\bullet) = -\ln(-\ln(\bullet))$. The value of Y, y, equals $(x-u)/a$ or $-\ln\left(-\ln\left(F_G(x)\right)\right)$; the T-year return period value of X, x_T equals $u - a\ln\left(-\ln\left(1-1/T\right)\right)$. The ascending ordered samples of X, denoted by $X_{i:n}$, i =1,…, n, are related to the ordered statistics of Y, $Y_{i:n}$, by

$$Y_{i:n} = (X_{i:n}-u)/a \tag{3}$$

Let $F_{Y_{i:n}}(y_{i:n})$ denote probability distribution function of $Y_{i:n}$. Its mean $_{\text{Ei}}$ and its median $_{\text{Mi}}$ are given by (Blom 1958, David and Nagaraja 2003)

$$\pi_{Ei} = i/(n+1) \tag{4}$$

and $\pi_{Mi} = F_{Beta}^{-1}\left(0.5; i, n-i+1\right)$, where $F_{Beta}^{-1}\left(\bullet; i, n-i+1\right)$ denotes the inverse beta distribution function with parameters i and $n - i$ +1. π_{Ei} is known as the Weibull plotting position, and π_{Mi} is used as the basis for Benard and Bos-Levenback (1953), and Filliben (1975).

The evaluation of the mean of $Y_{\text{i:n}}$, $\alpha_{i:n} = E(Y_{i:n})$ and the covariance of $Y_{\text{i:n}}$ and $Y_{\text{j:n}}$, $\beta_{i,j:n} = E(Y_{i:n}Y_{j:n}) - \alpha_{i:n}\alpha_{j:n}$ were discussed in Lieblein (1953, 1974), and in Balakrishnan and Chan (1992). Analytical solution is available only for $\alpha_{n:n}$ that is given by

$$\alpha_{n:n} = \gamma + \ln n \tag{5}$$

where the Euler's constant γ equals 0.5772. For other cases, it has been shown (Balakrishnan and Chan 1992, Hong *et al*. 2013) that the use of numerical integration can provide accurate results for samples size, at least, up to hundreds.

$\alpha_{\text{i:n}}$ could be approximated by $F_Y^{-1}(\pi_{Ei})$, but better approximation is obtained by (Blom 1958)

$$\alpha_{i:n} \approx F_Y^{-1}(\pi_{Bi}) \tag{6a}$$

where

$$\pi_{Bi} = (i-\alpha_n)/(n+1-\alpha_n-\beta_n) \tag{6b}$$

in which α_{n} and β_{n} are the correction factors to be determined. Note that the subscript n for the factors in Eq. (6(b)) emphasizes that they could be functions of the sample size.

By assuming the symmetry in the probability distribution (i.e., the consideration of $\pi_{Bi} = 1 - \pi_{Bn+1-i}$ and the application of Eq. (6(b)) result in $\alpha_{\text{n}} = \beta_{\text{n}}$) and by matching $\alpha_{n:n}$ given in Eq. (5) through the application of Eqs. (6(a)) and (6(b)), Gringorten (1963) recommended (i – 0.44)/ (n+0.12) as the plotting position (for $n \geq 20$). By relaxing the symmetry assumption and

considering that the correction factors depend on n, Cook and Harris (2013) estimated α_n and β_n based on the weighted least-squares fits but guided by the analysis procedure termed as the "Gringorten extended" methodology and asymptotic behaviour. They suggested $\alpha_n = 0.439 - 0.466/\ln n$ and $\beta_n = 0.448$, resulting in the following plotting position, π_{CHi}

$$\pi_{CHi} = (i - 0.439 + 0.466/\ln n)/(n + 0.113 + 0.466/\ln n) \tag{7}$$

A plotting position by considering the mean of order statistics for Gumbel variate was proposed by Pirouzi Fard (2010). The proposed plotting position does not consider the dependency of α_n and β_n on the sample size. Analysis was carried out in Hong (2013) to suggest a new plotting position (see Case 1 in Table 1) but by fitting $\alpha_{i:n}$, i =1,…,n, obtained using numerical integration for n within 3 to 300. Although the fit is excellent and its use for distribution fitting is practically identical to that by using $\alpha_{i:n}$ in the OL method, a drawback is that the suggested plotting position is less amenable than Eq. (7). To overcome this and motivated by the plotting position format recommended by Cook and Harris (2013), we carried out the following analysis to find simple suitable correction factors by:

1) Minimize the sum of the squared value of $e_{i:n}$, $e_{i:n} = \alpha_{i:n} - F_Y^{-1}(\pi_{Bi})$, to find values of α_n and β_n, denoted by $\hat{\alpha}_n$ and $\hat{\beta}_n$, for each considered n value; and

2) Select models for α_n and β_n, by try and error, and find their model parameters by minimizing $E_\alpha = \sum(\alpha_n - \hat{\alpha}_n)^2$ and $E_\beta = \sum(\beta_n - \hat{\beta}_n)^2$.

For the numerical analysis, values of $\hat{\alpha}_n$ and $\hat{\beta}_n$ are calculated for n equal to 3 to 1000 with an increment of 1 (i.e., 3(1)1000). The estimated values are shown in Fig. 1. A set of results was obtained by excluding $_{n:n}$, while the other set by including $_{n:n}$. This is because the closed form solution for $_{n:n}$ is available (see Eq. (5)) and could be used directly. The results presented in Fig. 1 indicate that $\hat{\alpha}_n$ and $\hat{\beta}_n$ are very smooth function of n, and are suitable to develop empirical models.

For the minimization of E and E defined previously, we considered the values of $\hat{\alpha}_n$ and $\hat{\beta}_n$ for n equal to 5(1)200 because often the sample size is small in practical extreme value analysis. Several empirical models tried for α_n and β_n and their corresponding results are included in Table 1. Values of E and E shown in the table indicate that Cases 2 and 6 provide excellent fit to $\hat{\alpha}_n$ and $\hat{\beta}_n$, while Cases 7, 8 and 9 are simpler to use. Based on the results shown in the table and for simplicity, we recommend the use of $\alpha_n = 0.37 - 0.232/\sqrt{n}$ and $\beta_n = 0.486$ for the plotting position (see Case 4 in Table 1)

$$\pi_{Bi} \approx \begin{cases} (i - 0.37 + 0.232/\sqrt{n})/(n + 0.144 + 0.232/\sqrt{n}), & i = 1,...,n-1 \\ \exp(-\exp(-0.5772)/n), & i = n \end{cases} \tag{8}$$

A verification analysis shows that for n = 500 and 1000, $\hat{\beta}_n$ equal to 0.486 is adequate and the

relative difference between $\alpha_n = 0.37 - 0.232/\sqrt{n}$ and $\hat{\alpha}_n$ is less than 4%.

Note that if the functional form used by Cook and Harris (2013) for α_n is considered, we obtained $\alpha_n = 0.394 - 0.223/\ln n$ resulting in (see Case 5 in Table 1)

$$\pi_{Bi} = \begin{cases} (i - 0.394 + 0.223/\ln n)/(n + 0.12 + 0.223/\ln n), & i = 1,...,n-1 \\ \exp(-\exp(-0.5772)/n), & i = n \end{cases} \quad (9)$$

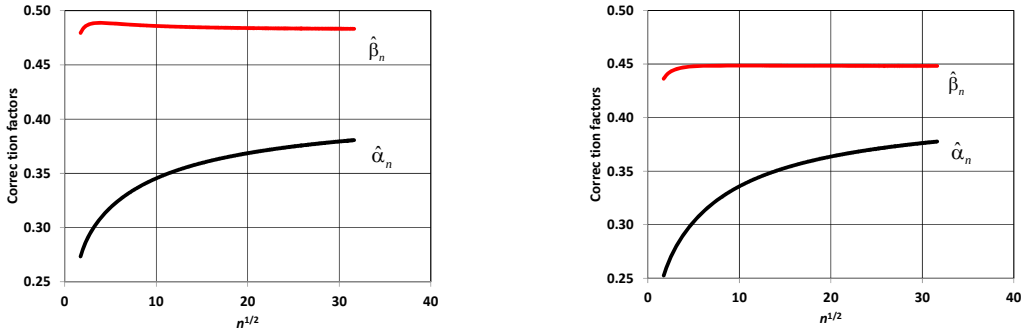
In terms of minimizing E_α, α_n for Eq. (8) is preferable than that for Eq. (9) (see Table 1).

It must be emphasized that while Eq. (7) is developed by considering large *n* and asymptotic behaviour, the recommended plotting position in Eq. (8) and the equations in Table 1 are developed for sample sizes often encountered in practical extreme value analysis.

Table 1 Suggested plotting position (analysis carried out based n = 5(1)200, unless otherwise indicated)

Case	Model and estimated model parameters	Error
1	$\alpha_n = -0.00054(\ln n)^2 + 0.0235\ln n + 0.2487$, $\beta_n = -0.001(\ln n)^2 + 0.0063\ln n + 0.4781$, Analysis was carried out using *n* = 3(1)100,150(50)300 given in Hong (2013)	
2	$\alpha_n = c_1 + c_2/n^{c_3}$; c_1=0.6860, c_2=-0.4376, c_3=0.0543; $\beta_n = d_1 + d_2/n^{d_3}$; d_1=0.4892, d_2=-0.0002, d_3=-0.5887	E_α=2.59E-6, E_β=2.60E-5
3	$\alpha_n = c_1 + c_2/\ln n$; c_1=0.3943, c_2=-0.2230; $\beta_n = d_1 + d_2/\ln n$; d_1=0.4831, d_2=0.0127	E_α =4.35E-3, E_β =1.28E-4
4	$\alpha_n = c_1 + c_2/\sqrt{n}$; c_1=0.3695, c_2=-0.2328; $\beta_n = d_1$; d_1=0.4862. Its simplified version $\alpha_n = 0.370 - 0.232/\sqrt{n}$; $\beta_n = 0.486$	E_α =2.67E-3, E_β =2.69E-4
5	$\alpha_n = c_1 + c_2/\ln n$; c_1=0.3943, c_2=-0.2230; β_n=d_1; d_1=0.4862 Its simplified version $\alpha_n = 0.394 - 0.223/\sqrt{n}$; $\beta_n = 0.486$	E_α =4.35E-3, E_β =2.69E-4
6	$\alpha_n = c_1 + c_2/n^{C_3}$; c_1=0.8493, c_2=-0.6294, c_3=0.0441; $\beta_n = d_1 + d_2/n^{d_3}$; d_1=0.4486, d_2=-0.0731, d_3=1.4053	E_α =1.27E-5, E_β =5.17E-6
7	$\alpha_n = c_1 + c_2/\ln n$; c_1=0.3950,c_2=-0.2701; $\beta_n = d_1$; d_1 = 0.4482	E_α =6.72E-3, E_β =1.63E-4
8	$\alpha_n = c_1 + c_2/\sqrt{n}$; c_1= 0.3650,c_2= -0.2823; $\beta_n = d_1$; d_1 = 0.4482	E_α =4.15E-3, E_β =1.63E-4
9	$\alpha_n = 0.439 - 0.466/\ln n$ and $\beta_n = 0.448$ given by Cook and Harris (2013)	

Note: The plotting position for Cases 1 to 5 is based on Eq. (6) but $\pi_{Bi} = \exp(-\exp(-0.5772)/n)$ when *i* = *n*. The plotting position for Cases 6 to 9 is based on Eq. (6)



(a) Excluding $\alpha_{n:n}$ in fitting Eq. (6(b)) (b) Including $\alpha_{n:n}$ in fitting Eq. (6(b))

Fig. 1 Variation of $\hat{\alpha}_n$ and $\hat{\beta}_n$ versus n

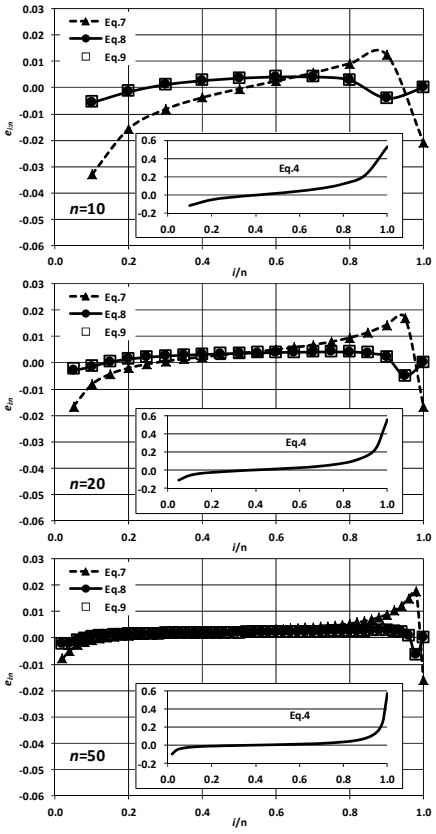


Fig. 2 $e_{i:n}$ for different plotting positions and n values

Since the minimum values of E and E alone cannot be used as the measure to indicate the accuracy of Eq. (6a), the difference $e_{i:n}$ by using the suggested plotting position (i.e., Eq. (8)) is calculated. The obtained values are presented in Fig. 2 for n = 10, 20 and 50. The figure shows that the absolute value of the error is less than 1%. The calculation of the difference is repeated by replacing the plotting position shown in Eq. (8) with the plotting positions presented in Eqs. (4), (7) and (9). The calculated $e_{i:n}$ is compared in Fig. 2, indicating that the absolute value of $e_{i:n}$ for Eq. (9) is almost identical to that for Eq. (8). The use of Eq. (7) can lead to the absolute value of $e_{i:n}$ up to 3.5% for n = 10; this value reduces as n increases. If Eq. (4) is used, the maximum of the absolute value of $e_{i:n}$ is at least an order of magnitude greater than that by using the plotting position recommended in Eq. (8).

2.2 Approximation to variance and covariance of order statistics

The importance of considering the variance and/or covariance to estimate the return period values was stressed in Harris (1996) and Hong (2013). They showed that the RMSE of the return period value estimated by using the OL method is greater than that by using the WL and GL methods, especially if the coefficient of variation (cov) of the Gumbel variate is large. Since the WL and GL methods require the use of $\beta_{i,i:n}$ and the numerical evaluation of $\beta_{i,i:n}$ is involved (Lieblein 1953, Balakrishnan and Chan 1992), a simple to use approximation to $\beta_{i,i:n}$ is desirable.

Closed form solution for $\beta_{i,i:n}$ is available only for $i = j = n$, resulting in $\beta_{n,n:n} = \pi^2 / 6$. For other values of i and j, $\beta_{i,i:n}$ can be approximated by (Blom 1958, David and Nagaraja 2003)

$$\beta_{i,j:n} \approx \frac{1}{\pi_{Ei}\pi_{Ej}\ln(\pi_{Ei})\ln(\pi_{Ej})} C\left(p_i, p_j\right) \tag{10a}$$

where $p_i = F_Y(y_{i:n})$, $C(\bullet,\bullet)$ denotes the covariance of its arguments and is given by,

$$C\left(p_i, p_j\right) = \frac{i(n+1-j)}{(n+1)^2(n+2)}, \quad \text{for } i \le j \tag{10b}$$

Substituting, Eq. (10(b)) and $\pi_{Ei} = i/(n+1)$ into Eq. (10(a)) results in,

$$\beta_{i,j:n} \approx \frac{1}{\ln(i/(n+1))\ln(j/(n+1))} \frac{(n+1-j)}{j(n+2)}, \quad \text{for } i \le j \tag{11}$$

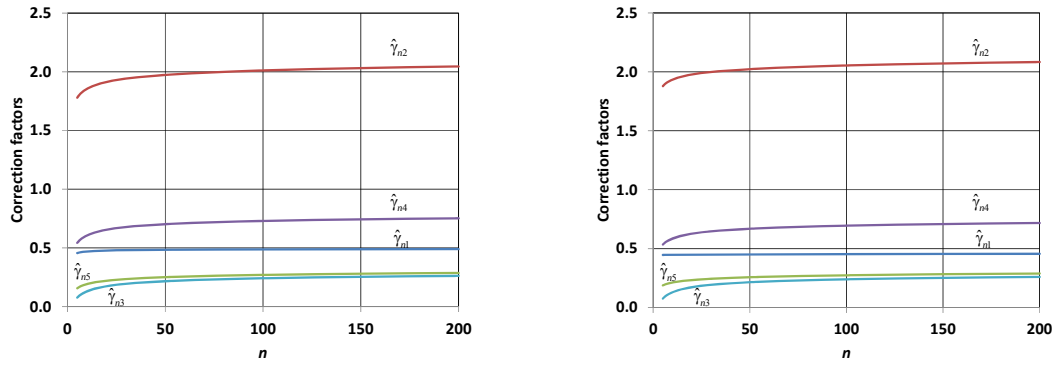
Pirouzi Fard and Holmquist (2008) noted that this approximation (i.e., its corresponding approximation for Gumbel minimum distribution) can be improved by considering additional correction factors

$$\beta_{i,j:n} \approx \begin{cases} \pi^2/6, & \text{for } i=j=n \\ \dfrac{n+1-j-\gamma_{n1}}{(n+2-\gamma_{n2})(j-\gamma_{n3})\ln\left(\dfrac{i-\gamma_{n5}}{n+1-\gamma_{n4}}\right)\ln\left(\dfrac{j-\gamma_{n3}}{n+1-\gamma_{n4}}\right)}, & \text{for } i \le j \text{ and } i \ne n \end{cases} \quad (12a)$$

where γ_{nk}, $k = 1,\ldots,5$, are the correction factors to be determined based on regression analysis. An alternative to Eq. (12(a)) is

$$\beta_{i,j:n} \approx \frac{n+1-j-\gamma_{n1}}{(n+2-\gamma_{n2})(j-\gamma_{n3})\ln\left(\dfrac{i-\gamma_{n5}}{n+1-\gamma_{n4}}\right)\ln\left(\dfrac{j-\gamma_{n3}}{n+1-\gamma_{n4}}\right)}, \quad \text{for } i \le j \quad (12b)$$

By fitting $\beta_{i,i:n}$, which are calculated using numerical procedure (Balakrishnan and Chan 1992, Hong *et al*. 2013), and by considering the above models, values of γ_{nk} for n=5(1)200, $\hat{\gamma}_{nk}$, are estimated based on weighted least-squares fits and are shown in Fig. 3. The results indicate that in all cases the correction factors do not change very much for n greater than about 25. Such smoothed trends are not shared by the results shown in Table 1 in Pirouzi Fard and Holmquist (2008). We suppose this is caused by the inaccuracy of $\beta_{i,i:n}$ values used in their calculation that are estimated by simulation. Inspection of the quality of fit for n = 5(1)200 (not shown in here to save space) indicates that Eq. (12(a)) provides better fit than Eq. (12(b)). Since the variations of values of $\hat{\gamma}_{n1}$ and $\hat{\gamma}_{n2}$ are small, simplified models to approximate $\beta_{i,i:n}$ are obtained by assigning $\gamma_{n1} = 0.5$ and $\gamma_{n2} = 2$ in Eqs. (12(a)) and (12(b)). The resulting models are listed in Table 2 as Cases 3 and 4. Inspection of the fitting for different n indicates Eq. (12(a)) is the most preferred, and equations for Cases 1 and 2 perform better than those for Cases 3 and 4. The equation for Case 4 to approximate $\beta_{i,i:n}$ performs the worst.



(a) Considering Eq. (12(a)) (b) Considering Eq. (12(b))

Fig. 3 Variation of $\hat{\gamma}_{nk}$ versus n

Table 2 Some suggested models to approximate variance and covariance (analysis carried out based n = 5(1)200)

Case	Plotting position	Model and estimated model parameters	Error
1	Eq. (12(a))	$\gamma_{nk} = c_{1k} + c_{2k} \times n^{c_{3k}}, k = 1,...,5$ c_{11} = -3.9244, c_{21}= 4.3825, c_{31}= 0.0014; c_{21} = 2.2145, c_{22}= -0.6407, c_{32}= -0.2505; c_{31} = 0.3401, c_{32}= -0.4446, c_{33}= -0.3274; c_{41} = 0.8324, c_{42}= -0.4986, c_{43}= -0.3441; c_{51} = 0.3783, c_{52}= -0.3214, c_{53}= -0.2368	E_1 = 5.13E-4, E_2 = 2.10E-4, E_3 = 1.15E-5, E_4 = 3.08E-5, E_5 = 1.88E-6
2	Eq. (12(a)) but $\gamma_{n1} = 0.5$ and $\gamma_{n2} = 2$	$\gamma_{nk} = c_{1k} + c_{2k} \times n^{c_{3k}}, k = 3,...,5$ c_{31} = 0.3507, c_{32}= -0.4691, c_{33}= -0.3058; c_{41} = 0.8544, c_{42}= -0.4595, c_{43}= -0.2959; c_{51} = 0.1299, c_{52}= 0.0663, c_{53}= 0.1597	E_3 = 1.13E-4, E_4 = 9.57E-5, E_5 = 4.74E-5
3	Eq. (12(b))	$\gamma_{nk} = c_{1k} + c_{2k} \times n^{c_{3k}}, k = 1,...,5$ c_{11} = 0.4444, c_{21}= 0.0012, c_{31}= 0.4420; c_{21} = 2.2902, c_{22}= -0.5390, c_{32}= -0.1797; c_{31} = 0.3379, c_{32}= -0.4383, c_{33}= -0.3262; c_{41} = 0.8284, c_{42}= -0.4420, c_{43}= -0.2612; c_{51} = 0.4434, c_{52}= -0.3141, c_{53}= -0.1343	E_1 = 4.43E-6, E_2 = 2.39E-4, E_3 = 1.73E-5, E_4 = 3.02E-5, E_5 = 8.84E-6
4	Eq. (12(b)) but $\gamma_{n1} = 0.5$ and $\gamma_{n2} = 2$	$\gamma_{nk} = c_{1k} + c_{2k} \times n^{c_{3k}}, k = 3,...,5$ c_{31} = 0.4083, c_{32}= -0.4776, c_{33}= -0.1868; c_{41} = 0.9128, c_{42}= -0.4740, c_{43}= -0.1816; c_{51} = 0.1054, c_{52}= 0.0813, c_{53}= 0.1485	E_3 = 1.07E-6, E_4 = 2.17E-6, E_5 = 1.81E-4

Models to fit $\hat{\gamma}_{n2}$ are selected by try and error, and the model parameters are determined by minimising $E_{\gamma k} = \sum \left(\gamma_{nk} - \hat{\gamma}_{nk} \right)^2$, k = 1,…,5. Some of the empirical models are presented in Table 2. Based on the sum of the squared error E_{k} alone, the model corresponds to Case 1 shown in Table 2 is preferred. However, for simplicity, Case 3 may be considered. An illustration of the difference by using Case 1 and Case 3 to approximate $\beta_{i,j:n}$ in terms of relative error is presented in Fig. 4. The figure shows that the use of Eq. (12(a)) leads to a relative error within -1% to 1% which is excellent.

If the use of the GL method (Lloyd 1952, Lieblein 1974, Hong *et al*. 2013) for the distribution fitting with approximate $\alpha_{i:n}$ and $\beta_{i,j:n}$ is considered, it is suggested that equations for Case 2 (or Case 1) shown in Table 1 and equations for Case 1 shown in Table 2 are to be employed.

2.3 Relation to Weibull distribution

The plotting position and the approximation to variance and covariance of the order statistics for the Gumbel distribution can be applied to other extreme value distributions, which can be transformed into the Gumbel (maximum) distribution form shown in Eq. (1) (David and Nagaraja 2003, Hong 2013). In particular, if the following Weibull distribution for the random variable *Z*,

which is often used to fit the (point-in-time) wind speed data (Holmes 2001, Hong 1994), is considered

$$F_W(z) = 1 - \exp\left(-\left(z / u_W\right)^{\alpha_W}\right) \tag{13}$$

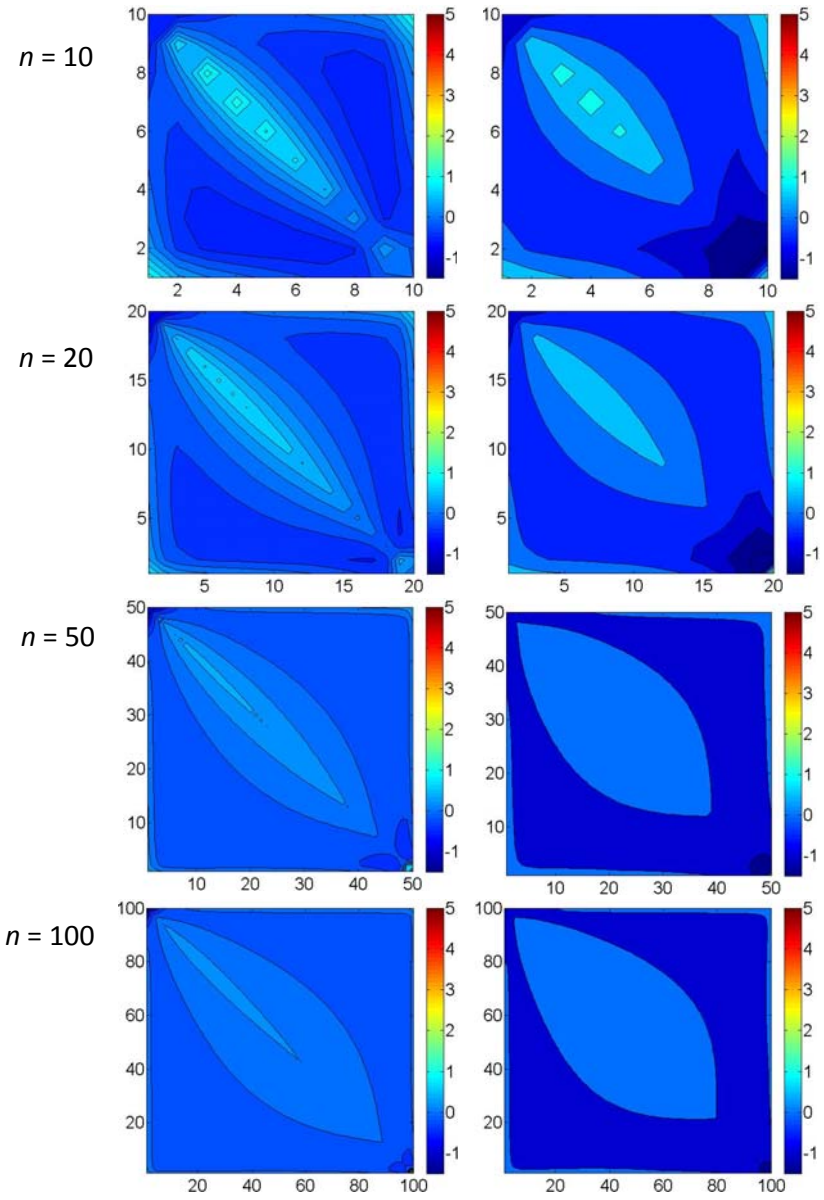


Fig. 4 Relative error (%) (i.e., (difference between approximation to $\beta_{i,j:n}$ - $\beta_{i,j:n}$)/ $\beta_{i,j:n}$) in approximating the variance and covaraince: Left column for approximation using Eq. (12(a)) (see Case 1 in Table 2) and right column for approximation using Eq. (12(b)) (see Case 3 in Table 2)

It can be shown that –ln(z) is Gumbel distributed with the location and scale parameters (u, a) = (-lnu_W, 1/1/α_W). Therefore, the plotting positions and the variance and covariance approximations developed in the previous section can be directly applied to –ln(z). Alternatively, if the use of ln(z) rather than –ln(z) is preferred, the plotting position for the i-th order of samples from the Weibull distribution, $\left(\pi_{Bi}\right)_{Weibull}$ is equal to 1 minus the (n+1-i)-th order of the samples from the Gumbel distribution. That is

$$\left(\pi_{Bi}\right)_{Weibull} = (i-\beta_n)/(n+1-\alpha_n-\beta_n) \tag{14}$$

where α_n and β_n on the right hand side of the equation are those derived for the Gumbel distribution and discussed in the previous section. The relations between the moments of the order statistics of the Weibull variate to those of the Gumbel variate are already given in Lieblein and Zelen (1956).

Because of the above, in the following we only carry out the numerical assessment of the bias and RMSE considering the Gumbel distribution, as the conclusions are directly applicable to the transformed Weibull variate (i.e., -lnz).

3. Use of the plotting position in distribution fitting and quantile estimation

3.1 Bias and RMSE of distribution parameters and quantile estimators

Consider that ζ represents the distribution parameter a or u, or x_T (i.e., quantile with exceedance probability of 1/T). The performance of the estimator of ζ, $\hat{\zeta}$, can be judged using the relative bias and RMSE

$$Bias = \left(\sum_{i=1}^{N}\left((\hat{\zeta}_i-\zeta)/\zeta\right)\right)/N \tag{15a}$$

and

$$RMSE = \left(\frac{1}{N}\sum_{i=1}^{N}\left((\hat{\zeta}_i-\zeta)/\zeta\right)^2\right)^{1/2} \tag{15b}$$

where N is the number of replicas and, $\hat{\zeta}_i$ denotes the estimator of ζ for the i-th replica.

Numerical estimation of the relative bias and RMSE for the OL method (using $\alpha_{i:n}$ or an adopted plotting position) is carried out below using the simulation technique. For each replication, the simulation procedure generates n samples from the Gumbel distribution for given distribution parameters u and a; the ordered samples $x_{i:n}$ are used to estimate

$$(\hat{a},\ \hat{u}) = (s_{x\alpha}/s_{\alpha}^2,\ m_x-\hat{a}m_{\alpha}) \tag{16a}$$

and

$$\hat{x}_T = \hat{u}-\hat{a}\ln\left(-\ln\left(1-1/T\right)\right) \tag{16b}$$

where $s_\alpha^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\alpha_{i:n}^2\right) - m_\alpha^2$, $s_{x\alpha} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{i:n}\alpha_{i:n}\right) - m_x m_\alpha$, $m_x = \frac{1}{n}\sum_{i=1}^{n} x_{i:n}$, and $m_\alpha = \frac{1}{n}\sum_{i=1}^{n}\alpha_{i:n}$.

If the plotting position π_i is adopted to approximate $\alpha_{i:n}$, $\alpha_{i:n}$ in Eq. (16) is replaced by $-\ln(-\ln(\pi_i))$. The replication is repeated *N* times and values of $(\hat{a},\hat{u},\hat{x}_T)$ in Eqs. (16(a)) and (16(b)) are calculated.

The parametric analysis can be simplified by noting that the positive scaling of both model parameters (*u*, *a*) for the Gumbel distribution does not affect the relative bias and RMSE of the estimators. Therefore, without loss of generality, *a* can be assigned equal to one and varying *u* for parametric study. Note that in such a case, *u* is related to the cov of *X*, v_x, by $u = \pi/\left(v_X\sqrt{6}\right) - 0.5772$; *u* equals 25.07, 12.25, 7.97, 5.84, 3.70, 2.63, 1.99, 1.56, 1.26, 1.06, 0.85 and 0 for v_X equal to 0.05(0.05)0.20, 0.3(0.1)0.9 and 2.22, respectively.

For a few selected *n* and *u* = 0, the relative *Bias* and *RMSE* of the estimators of (a, u, x_T), $(\hat{a},\hat{u},\hat{x}_T)$, obtained by using Eqs. (15) and (16) are presented in Table 3 for the plotting positions given in Eqs. (4), (7), (8) and (9). For the analysis, *N* equal to 100,000 is used because for *N* greater than 100,000, the changes in *RMSE* are negligible, and the changes in bias as compared to its standard deviation are very small. For comparison purpose, Table 3 also includes the results by using the OL, WL and GL methods (with accurate $\alpha_{i:n}$ and $\beta_{i,j:n}$). The results for OL, WL and GL methods are practically identical to those presented in Hong *et al*. (2013) and Hong (2013); for consistency they are re-calculated using the same sets of simulated samples for the same number of replications in this study. The table indicates that:

1) The relative Bias and RMSE of $(\hat{a},\hat{u},\hat{x}_T)$ by using the OL method with the plotting position shown in Eq. (8) are almost identical to those by using OL method with (accurate) $\alpha_{i:n}$. This indicates that the recommended plotting position is adequate and that its use provides good approximation to $\alpha_{i:n}$.
2) The magnitudes of the relative Bias and RMSE of $(\hat{a},\hat{u},\hat{x}_T)$ by using the OL method with Eq. (8) are less than or equal to those obtained by considering Eqs. (4) and (7). The differences between using Eq. (8) and Eq. (9) are negligible.
3) The relative RMSE by using OL method with $\alpha_{i:n}$ or with an adopted plotting position is greater than those obtained by the WL and GL methods. This indicates that the use of the WL and GL methods are preferred, an observation already made in other studies (Harris 1996, Hong 2013).
4) The use of Eq. (4) is inadequate since the objective of the fitting is to estimate quantile free from bias and with minimum RMSE. This observation is consistent with the observation made by others, including Cunnane (1978), Harris (1996) and Fuglem *et al*. (2013).

To see the effect of the cov of the Gumbel variate on the relative bias and RMSE, we repeat the above analysis but for *n* equal to 20 only. The obtained results are shown in Fig. 5, indicating again that the relative Bias and RMSE obtained by using the OL method with the plotting position presented in Eq. (8) are practically identical to those obtained by using OL method with $\alpha_{i:n}$. The figure also shows that the relative bias and RMSE decrease largely as the *u* increases (i.e., as the cov decreases).

Table 3 Relative bias and RMSE of the estimators of (a, u, x_T), $(\hat{a}, \hat{u}, \hat{x}_T)$, for different methods (see Eqs. (15(a)) and (15(b)), except for the estimator of u, values are calculated without dividing by u since $u = 0$ is considered)

Parameter		Relative Bias(×100)							Relative RMSE(×10)						
		Eq.8	Eq.9	Eq.7	Eq.4	OL	WL	GL	Eq.8	Eq.9	Eq.7	Eq.4	OL	WL	GL
n=20	a	0.01	0.00	0.05	11.46	0.01	-0.01	-0.01	2.23	2.23	2.24	2.70	2.23	1.89	1.82
	u	0.18	0.20	0.71	-0.65	-0.02	-0.01	-0.01	2.40	2.40	2.40	2.39	2.40	2.39	2.36
	x_{30}	0.06	0.06	0.26	11.27	0.00	-0.02	-0.01	2.43	2.43	2.44	2.87	2.43	2.18	2.12
	x_{50}	0.06	0.05	0.23	11.29	0.00	-0.02	-0.01	2.39	2.39	2.40	2.84	2.39	2.13	2.07
	x_{100}	0.05	0.04	0.21	11.32	0.00	-0.01	-0.01	2.36	2.36	2.37	2.81	2.36	2.08	2.02
	x_{500}	0.04	0.03	0.17	11.35	0.01	-0.01	-0.01	2.32	2.32	2.32	2.77	2.32	2.02	1.96
n=30	a	0.03	0.03	0.09	8.97	0.03	0.02	0.02	1.83	1.83	1.84	2.16	1.83	1.54	1.47
	u	0.17	0.18	0.63	-0.70	-0.01	-0.01	-0.01	1.96	1.96	1.96	1.95	1.96	1.95	1.93
	x_{30}	0.08	0.08	0.27	8.76	0.02	0.01	0.02	1.99	1.99	2.00	2.30	1.99	1.78	1.72
	x_{50}	0.08	0.07	0.25	8.79	0.02	0.02	0.02	1.96	1.96	1.97	2.27	1.96	1.73	1.68
	x_{100}	0.07	0.07	0.23	8.82	0.03	0.02	0.02	1.94	1.94	1.94	2.25	1.93	1.70	1.64
	x_{500}	0.06	0.06	0.19	8.85	0.03	0.02	0.02	1.90	1.90	1.90	2.22	1.90	1.65	1.58
n=50	a	0.05	0.05	0.11	6.53	0.04	0.02	0.02	1.43	1.43	1.43	1.63	1.43	1.19	1.12
	u	0.14	0.14	0.48	-0.69	0.00	0.00	0.00	1.52	1.52	1.52	1.51	1.52	1.51	1.49
	x_{30}	0.09	0.09	0.26	6.32	0.04	0.02	0.02	1.55	1.55	1.55	1.74	1.55	1.38	1.32
	x_{50}	0.08	0.08	0.24	6.35	0.04	0.02	0.02	1.53	1.53	1.53	1.72	1.53	1.34	1.29
	x_{100}	0.08	0.08	0.22	6.37	0.04	0.02	0.02	1.51	1.51	1.51	1.70	1.51	1.31	1.26
	x_{500}	0.07	0.07	0.19	6.41	0.04	0.02	0.02	1.48	1.48	1.48	1.68	1.48	1.28	1.22
n=100	a	0.03	0.03	0.09	4.15	0.01	0.01	0.00	1.02	1.02	1.02	1.13	1.02	0.84	0.79
	u	0.08	0.08	0.29	-0.60	0.00	0.00	0.00	1.07	1.07	1.07	1.07	1.07	1.07	1.05
	x_{30}	0.05	0.05	0.18	3.97	0.01	0.00	0.00	1.10	1.10	1.10	1.20	1.10	0.97	0.93
	x_{50}	0.05	0.05	0.17	3.99	0.01	0.00	0.00	1.09	1.09	1.09	1.18	1.09	0.95	0.91
	x_{100}	0.05	0.05	0.15	4.02	0.01	0.00	0.00	1.07	1.07	1.07	1.17	1.07	0.93	0.88
	x_{500}	0.04	0.04	0.14	4.05	0.01	0.00	0.00	1.05	1.05	1.06	1.16	1.05	0.90	0.85

Also, the estimation of the relative Bias and RMSE of the estimators of (a, u, x_T), $(\hat{a},\hat{u},\hat{x}_T)$, such as those shown in Table 3 and Fig. 5 is carried out by using the WL and GL methods (Lloyd 1952, Lieblein 1974, Harris 1996, Hong *et al*. 2013) but with the approximate moments of order statistics $\alpha_{i:n}$ and $\beta_{i,j:n}$ given in Table 1 for Case 2 (or 1) and in Table 2 for Case 1. In all cases, the relative RMSE obtained by using the approximate $\alpha_{i:n}$ and $\beta_{i,j:n}$ values is identical to that shown in Table 3 and Fig. 5 for the corresponding distribution fitting method. The relative bias obtained by using the approximate $\alpha_{i:n}$ and $\beta_{i,j:n}$ in the WL and GL methods is greater than that obtained by using accurate $\alpha_{i:n}$ and $\beta_{i,j:n}$. The value of the relative bias is less than 0.25% but greater than 0. This indicates that a slight conservative but biased estimate of the return period values is obtained by using the approximate $\alpha_{i:n}$ and $\beta_{i,j:n}$ in the WL and GL method.

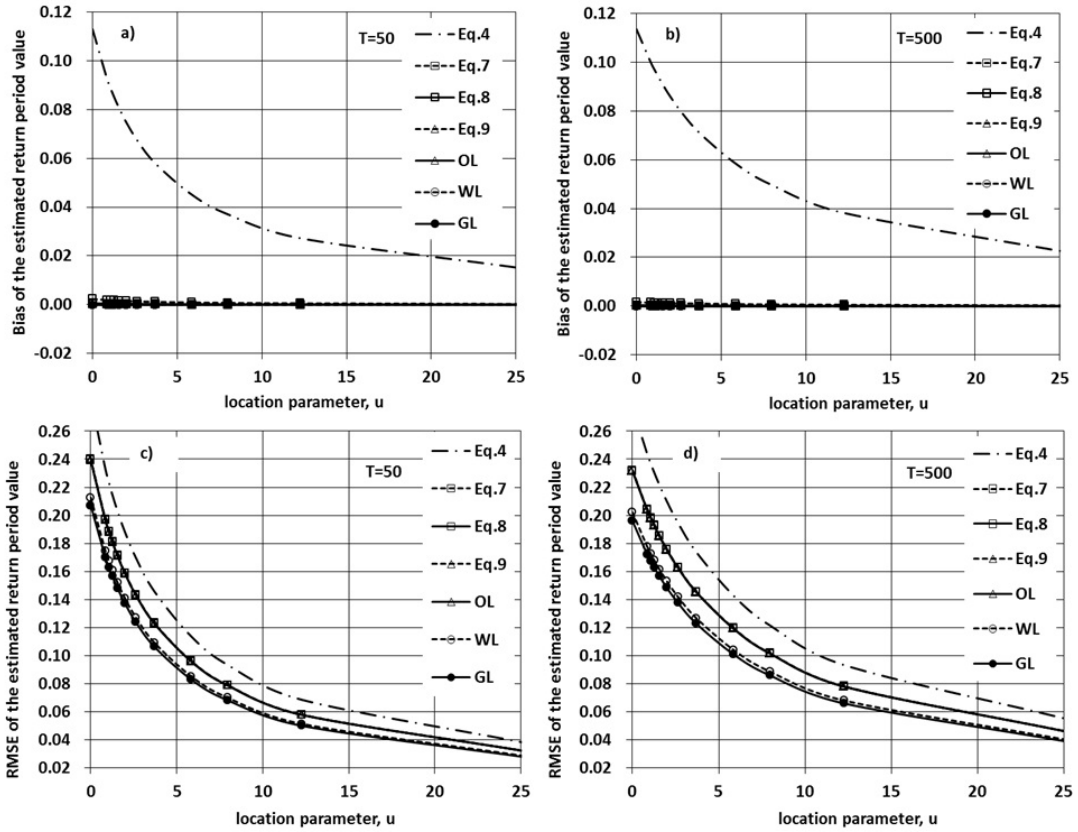


Fig. 5 Relative bias and RMSE of $\hat{x}_T$ by using different fitting method and plotting positions (T in year)

3.2 Illustrative applications

For practical illustrative application, we consider the problem of estimating the return period of annual maximum wind speed at seven meteorological stations in Canada shown in Table 4. The wind record processing and adjustment for the stations together with the estimated return period values of the annual maximum hourly-mean wind speed by using the WL and GL methods were presented in Hong *et al*. (2013). For the wind speed recorded at these stations, the use of Akaike information criterion indicates that the Gumble distribution for the annual maximum extreme wind speed is preferable than the generalized extreme value distribution.

By using the data sets, extreme value analysis using the OL method with $\alpha_{i:n}$ and the recommended plotting position shown in Eq. (8) are carried out using Eq. (16). The estimated 50-year and 500-year return period values are also shown in Table 4. The results indicate that the use of OL method with the recommended plotting position shown in Eq. (8) lead to the return period values that are practically the same as those obtained by using the OL method with $\alpha_{i:n}$. Also, analysis carried out by replacing Eq. (8) with Eq. (9) results in identical estimated return period values. The predicted return period values by using the Weibull plotting position (i.e., Eq. (4)) or any other plotting positions for the illustrative applications are not included to avoid potential misleading interpretation and conclusions.

As expected, the predicted return period values by the OL, WL and GL methods differ. The absolute value of the relative difference between the OL and GL methods is less than 3% for 50-year return period values, and less than 4% for 500-year return period value. This absolute value by considering WL and GL methods is less than 2% for 50-year return period values, and less than 3% for 500-year return period value.

The return period values estimated by using the GL (or WL) method with recommended approximation to $\alpha_{i:n}$ and $\beta_{i,j:n}$ are not shown because they are, up to one decimal point, identical to those for the GL (or WL) method shown in Table 4.

Table 4 Return period values of annual maximum hourly-mean wind speed (km/hr) for selected Canadian meteorological stations (The values for the WL and GL methods are taken from Hong *et al*. (2013). The results for the OL, WL and GL methods are obtained using accurate $\alpha_{i:n}$ and $\beta_{i,j:n}$)

Location	# of years	Method used for estimating 50-year return period value				Method used for estimating 500-year return period value			
		Eq.8	OL	WL	GL	Eq.8	OL	WL	GL
Edmonton Int'l A	50	81.41	81.40	80.98	81.58	94.99	94.98	94.26	95.25
Regina Int'l A	46	97.32	97.31	99.30	98.62	112.30	112.29	115.47	114.44
Ottawa Int'l A	50	86.70	86.69	89.66	89.29	100.74	100.73	105.50	105.01
Fredericton A	42	77.76	77.75	74.23	75.72	90.21	90.19	84.49	86.83
Halifax Int'l A	50	103.60	103.59	106.21	105.27	123.82	123.81	128.04	126.59
Charlottetown A	40	92.74	92.73	93.95	93.76	107.53	107.52	109.51	109.24
St. John's A	34	120.95	120.94	118.34	119.79	141.16	141.15	136.91	139.19

4. Conclusions

We presented a new simple to use plotting position for the Gumbel distribution. The plotting position depends on the sample size; its application can approximate well the mean of the order statistics; and its use with OL method provides the relative bias and RMSE error almost identical to those by using the accurate mean of the order statistics. Although the suggested plotting position is derived based on sample size up to 200, it was validated for sample size up to 1000.

We also recommended sets of equations, albeit less convenient to use, that can provide excellent approximation to the first two moments of order statistics for the Gumbel variate. These approximations can be used to facilitate the distribution fitting by using the WL and GL methods. It must be emphasized that the conclusions and error analysis for the recommended approximations are for samples size less than about 200, which is sufficient for most practical cases dealing with many practical extreme value analysis problems.

Acknowledgments

Financial support received from National Science and Engineering Research Council of Canada and the University of Western Ontario is much appreciated.

References

Balakrishnan, N. and Chan, P.S. (1992), "Order statistics from extreme value distribution, I tables of means, variances and covariances", *Commun. Stat. - Simul.*, **21**(4), 1199-1217.

Benard, A. and Bos-Levenbach, E.C. (1953), *Het uitzetten van waarnemingen op waarschijnlijkdeids-papier*, (The Plotting of Observations on Probability Paper), Statististica Neerlandica, **7**, 163-173.

Blom, G. (1958), *Statistical Estimates and Transformed Beta-Variables*, New York, John Wiley & Sons.

Castillo, E. (1988), *Extreme Value Theory in Engineering*, Academic Press, New York.

Cook, N.J. (2012), "Rebuttal of "Problems in the extreme value analysis", *Struct. Saf.*, **34**(1), 418-423.

Cook, N.J. and Harris, R.I. (2013), "The Gringorten estimator revisited", *Wind Struct.*, **16**(4), 355-372.

Cunnane, C. (1978), "Unbiased plotting-positions - a review", *J. Hydrology*, **37**, 205-222.

David, H.A. and Nagaraja, H.N. (2003), *Order Statistics*, 3rd Ed., John Wiley & Sons.

Filliben, J.J. (1975), "The probability plot correlation test for normality", *Technometrics*, **17**(1), 111-117.

Fuglem, M., Parr, G. and Jordaan, I.J. (2013), "Plotting positions for fitting distributions and extreme value analysis", *Can. J. Civil. Eng.*, **40**(2), 130-139.

Genschel, U. and Meeker, W.Q. (2010), "A comparison of maximum likelihood and median-rank regression for Weibull estimation", *Quality Eng.*, **22**(4), 236-255.

Goda, Y. (2011), "Plotting-position estimator for the L-moment method and quantile confidence interval for the GEV, GPA and Weibull distribution applied for extreme wave analysis", *Coast. Eng. J.*, **53**(2) 111-149.

Gringorten, I.I. (1963), "A plotting rule for extreme probability paper", *J. Geophys. Res.*, **68**, 813-814.

Harris, R.I. (1996), "Gumbel re-visited - a new look at extreme value statistics applied to wind speeds", *J. Wind Eng. Ind. Aerod.*, **59**, 1-22.

Harter, H.L. (1984), "Another look at plotting positions", *Commun. Stat. - Theor. M.*, **13**(13), 1613-1633.

Holmes, J.D. (2001), *Wind loading of structures*, Spon Press, New York.

Hong, H.P. (2013), "Selection of regressand for fitting the extreme value distributions using the ordinary, weighted and generalized least-squares methods", *Reliab. Eng. Syst. Safe.*, **118**, 71-80.

Hong, H.P. (1994), "A note on extremal analysis", *Struct. Safe.*, **13**(4), 227-233.
Hong, H.P., Li, S.H. and Mara, T. (2013), "Performance of the generalized least-squares method for the extreme value distribution in estimating quantiles of wind speeds", *J. Wind Eng. Ind. Aerod.*, **119**, 121-132.
Jordaan, I.J. (2005), *Decisions under uncertainty: probabilistic analysis for engineering decisions*, Cambridge University Press, New York.
Lieblein, J. (1953), "On the exact evaluation of the variances and covariances of order statistics in samples from the extreme-value distribution", *Annal. Math. Stat.*, **24**(2), 282-287.
Lieblein, J. (1974), *Efficient methods of extreme-value methodology*, report NBSIR 74-602, National Bureau of Standards, Washington.
Lieblein, J. and Zelen, M. (1956), "Statistical investigation of the fatigue life of deep-groove ball bearings", *J. Res. National Bureau Standards*, **57**(5), 273-316.
Lloyd, E.H. (1952), "Least–squares estimation of location and scale parameters using order statistics", *Biometrika*, **39**(1-2), 88-95.
Pirouzi Fard, M.N. and Holmquist, B. (2008), "Approximations of variances and covariances for order statistics from the standard extreme value distribution", *Commun. Stat. – Simul. C.*, **37**(8), 1500-1506.
Pirouzi Fard, M.N. (2010), "Probability plots and order statistics of the standard extreme value distribution", *Comput Stat.*, **25**(2), 257–267.

CC