DOI: http://dx.doi.org/10.12989/was.2013.16.4.355

# The Gringorten estimator revisited

Nicholas John Cook * and Raymond Ian Harris

*RWDI, Unit 4, Lawrence Way, Dunstable, Bedfordshire, LU6 1BD, UK*

*(Received March 20, 2012, Revised May 12, 2012, Accepted June 18, 2012)*

**Abstract.** The Gringorten estimator has been extensively used in extreme value analysis of wind speed records to obtain unbiased estimates of design wind speeds. This paper reviews the derivation of the Gringorten estimator for the mean plotting position of extremes drawn from parents of the exponential type and demonstrates how it eliminates most of the bias caused by the classical Weibull estimator. It is shown that the coefficients in the Gringorten estimator are the asymptotic values for infinite sample sizes, whereas the estimator is most often used for small sample sizes. The principles used by Gringorten are used to derive a new Consistent Linear Unbiased Estimator (CLUE) for the mean plotting positions for the Fisher Tippett Type 1, Exponential and Weibull distributions and for the associated standard deviations. Analytical and Bootstrap methods are used to calibrate the bias error in each of the estimators and to show that the CLUE are accurate to better than 1%.

**Keywords:** plotting position; linear unbiased estimators; extreme values; peak-over-threshold; weibull distribution; exponential distribution; Fisher Tippett Type 1 distribution; method of independent storms

## 1. Introduction

### *1.1. Expectation and standard error*

It is a truth universally acknowledged, that the best unbiased estimate of any random variate, *x*, is the ensemble average over an infinite number of samples, $\langle x\rangle$, usually called the “expectation”, and that the corresponding standard error of a single sample is the standard deviation, $\sigma(x)$ of the ensemble. Hence, for a set of *N* values drawn from a random variable, *x*, and ranked in ascending order, *m* = 1 ... *N*

$$x_{m:N} = \langle x_{m:N} \rangle + \varepsilon \tag{1}$$

where $x_{m:N}$ is the *m*-th ranked value out of *N* values (the *m*-th smallest value) and the difference, $\varepsilon$, between the value of each rank and the expectation for any single sample is $\pm\sigma(x_{m:N})$.

By definition, the cumulative probability distribution (CDF), *P*(*x*), is a single-valued invertible function of *x* which is uniformly distributed over the range $0 \le P \le 1$. The CDF is a “mapping” function in the sense that *x* uniquely defines *P*(*x*) and vice-versa, i.e., $x \Leftrightarrow P(x)$, so that any

∗Corresponding author, Dr., E-mail: nick.cook@rwdi.com function of *x*, *y*(*x*), can be evaluated in terms of *y*(*P*(*x*)). The expectation $\langle y(P_{m:N})\rangle = \langle y_{m:N}\rangle$ is obtained by evaluating the Binomial expression

$$\langle y_{m:N}\rangle = \frac{N!}{(m-1)!(N-m)!}\int_0^1 y(P_{m:N})P^{m-1}[1-P]^{N-m}\,dP \qquad (2)$$

and the corresponding standard deviation is given by evaluating

$$\sigma(y_{m:N}) = \sqrt{\langle y^2(P_{m:N})\rangle - \langle y(P_{m:N})\rangle^2} \qquad (3)$$

Because $x_{m:N}$ is a random variate, it follows that $P(x_{m:N}) \equiv P_{m:N}$ is also random so, from Eq. (1), the best unbiased estimate of $P_{m:N}$ is $\langle P_{m:N}\rangle$ and the standard error for a single sample is $\sigma(P_{m:N})$. This forms the core of the frequency interpretation of the probability, *P*, (Cramer, 1946) and is illustrated by Fig. 1 , which shows $P_{m:N}$ for samples of size $N = 9$ stabilizing onto $\langle P_{m:N}\rangle = m/10$ as the number of trials, *n*, increases. Note that the rate of this stabilisation is quite slow and that the errors associated with the single first sample are quite large.

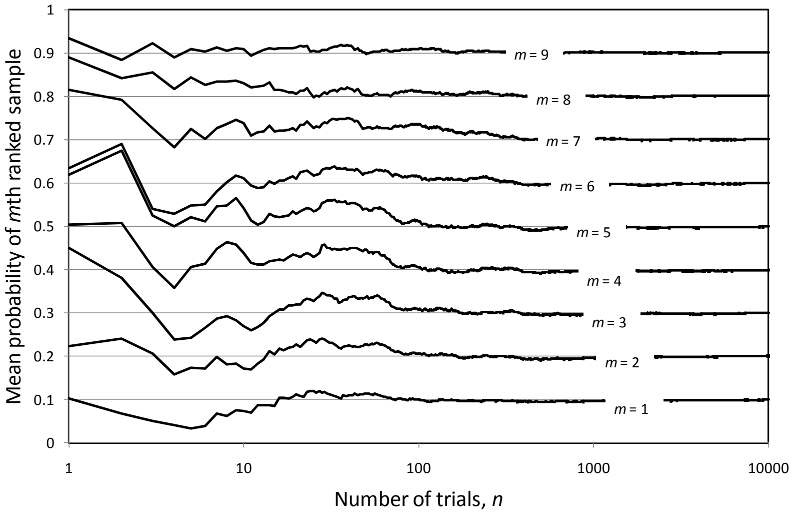


Fig. 1 Ensemble mean probability, $\langle P_{m:N}\rangle$, for each rank, *m*, of ranked samples of size *N* = 9 for *n* trials, accumulating as *n* increases

### *1.2 The Weibull estimator*

The ensemble mean probability $\langle P_{m:N}\rangle$ of the *m*-th smallest value in a ranked sample of size *N* for an infinite number of samples is given by evaluating Eq. (2) as

$$\langle P_{m:N} \rangle = \frac{m}{N+1} \tag{4}$$

(Castillo 1988, Eq. (2.41)). The corresponding standard deviation is given by

$$\sigma(P_{m:N}) = \sqrt{\frac{m(N-m+1)}{(N+1)^2(N+2)}} \tag{5}$$

(Castillo 1988, Eq. (2.42)).

The first use of Eq. (4) as the estimator for the CDF estimated from the order statistics of a single sample is usually attributed to Weibull (1939). The associated standard error given by Eq. (5) is rarely acknowledged in published analyses of wind engineering data, either directly or through the provision of confidence limits.

### *1.3 The Gringorten estimator*

While Eq. (4) gives the best unbiased estimate of the mean probability $\langle P_{m:N} \rangle$ corresponding to each rank, it does not give the best unbiased estimate of the corresponding value $\langle x_{m:N} \rangle$ unless the distribution $P(x)$ is a linear function of *x*. All probability distributions in nature are non-linear and tend to follow an S-shaped curve, so the inequality

$$\langle P(x) \rangle \neq P(\langle x \rangle) \tag{6}$$

generally applies. In most analysis and design applications, the aim is not to obtain the best unbiased estimate of probability for a given observational value, but is to obtain the best unbiased estimate of the variable for a datum (design) probability, i.e., the best estimate of a quantile. Eq. (6) shows that these two aims are not concomitant. The task is to find the plotting position that minimises bias in $P(\langle x \rangle)$.

The Gringorten (1963) estimator specifically addresses the estimation of the plotting position corresponding to $\langle x_{m:N} \rangle$ for ranked extremes drawn from parents of the exponential type. These fall into the domain of attraction of the asymptotic Fisher Tippett Type 1 (FT1) distribution, also known as the Gumbel distribution

$$P(x) = \exp\left(-\exp\left(-\frac{x-U}{b}\right)\right) \tag{7}$$

where the mode, *U*, (location parameter) and the dispersion, *b*, (scale parameter) are constants.

Estimating $\langle x \rangle$ for a FT1 distribution requires Eq. (7) to be rendered linear in terms of *x*, by taking logarithms twice, before applying the ensemble average operator, $\langle\ \rangle$. Hence

$$\langle x\rangle = b\times\langle -\ln(-\ln(P(x)))\rangle + U \tag{8}$$

The non-dimensional form of Eq. (8) is

$$\langle y\rangle = \langle -\ln(-\ln(P(x)))\rangle = \frac{\langle x\rangle - U}{b} \tag{9}$$

where $y$, the Gumbel (1958) “reduced variate”, is the non-dimensional form of $x$.

Gringorten (1963) derived an estimator for the probability of $\langle y_{m:N}\rangle$ in Eq. (9) of the form

$$P(\langle y_{m:N}\rangle) = \frac{m-A}{N+1-2A} \tag{10}$$

by using Gumbel’s expression for the top rank $\langle y_{N:N}\rangle = \ln(N)+\gamma$ (Gumbel 1958, p 116), where $\gamma$ = 0.57721... is Euler’s constant, and evaluating Eq. (10) to give $A$ = 0.43854, rounded to $A$ = 0.44. Appendix A shows that this is the asymptotic value obtained in the limit as $N \to \infty$, whereas the Gringorten estimator is typically applied to sample sizes that are quite small. Unfortunately, Gringorten (1963) did not derive an estimator for the corresponding standard deviation.

### *1.4 The non-parametric bootstrap*

The non-parametric bootstrap (Efron and Tibishirani 1993) is a Monte-Carlo sampling method that replicates the frequency interpretation of probability by generating random values in the range $0 \le P \le 1$ from a uniform random number generator. These random values of $P$ are mapped to the corresponding value of $x$ using an appropriate CDF, $P(x)$. Bootstrapping is an extremely useful technique for drawing inferences where a theoretical approach is not available or too difficult. Bootstrapping is used in wind engineering analyses to derive plotting positions, confidence limits and fitting weights (Naess and Clausen 2001, Cook 2004).

Fig. 2 shows the results for $N$ = 20, corresponding to a 20-year record of annual maxima, obtained by bootstrapping $10^4$ trials of $\langle y_{m:N}\rangle$ from the FT1 distribution, Eq. (9), plotted on the standard Gumbel axes at both the Weibull and the Gringorten plotting positions. The generating FT1 model is shown by the straight line, with unity slope and zero intercept. The Weibull estimator, Eq. (4), has a systematic bias which exaggerates the slope by 11%, while the Gringorten estimator replicates the model slope to within 1%.

Also shown in Fig. 2 are the 5% and 95% confidence limits for the data values and for the model fit[*], corresponding to the sampling error expected from a single sample. The bias produced by the Weibull estimator is smaller than the potential sampling error and is always conservative.

[*] The confidence limits for the data points were obtained by compiling the histogram of values for each rank from $10^4$ trials and taking the 500th value from the bottom (5%) and the 500th value from the top (95%). The confidence limits for the model fit were obtained by fitting each trial to the FT1 distribution using weighted least squares to determine the mode, $U$, and dispersion, $b$, then compiling the histograms of model values at a range of plotting positions and taking the 500th value from the bottom and the 500th value from the top.

The Gringorten estimator removes this bias almost entirely and closely replicates the generating FT1 model.

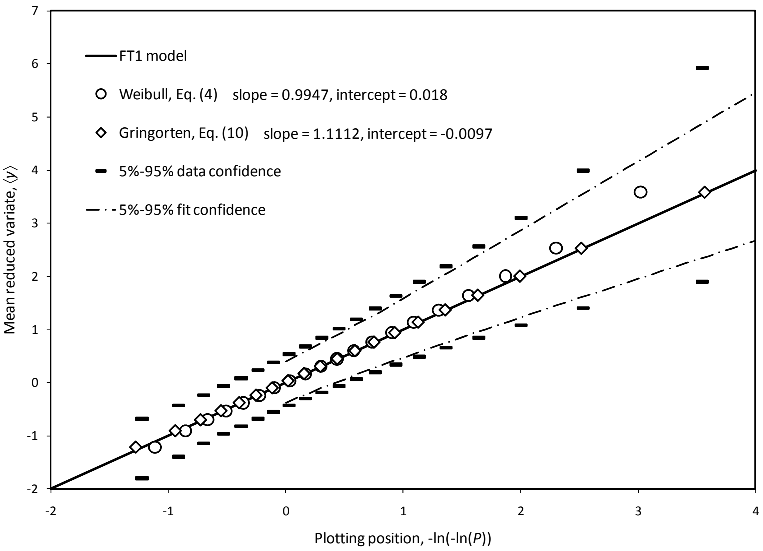


Fig. 2 Weibull and Gringorten estimators for $N = 20$ compared with $10^4$ bootstrap trials of FT1 distribution

## 2. Consistent linear unbiased estimator (CLUE)

### 2.1 General form of the CLUE

The Weibull and Gringorten estimators are linear estimators in the form of Eq. (10) with values of the coefficient, $A$, of $A = 0$ and $A = 0.44$, respectively, which apply for large sample sizes. A more general linear estimator which allows for different rates of convergence in the upper and lower tails is given by

$$P(\langle y_{m:N} \rangle) = \frac{m - A}{N + 1 - A - B} \tag{11}$$

This estimator applies consistently when $A$ and $B$ are made functions of the sample size, $N$. Whereas Gringorten solved Eq. (10) for one unknown coefficient using the expectation of the top rank, extending the Gringorten methodology for Eq. (11) requires a pair of simultaneous equations to be solved using the top and bottom ranks

$$A = \frac{N + 1 - B - 1/P_{1:N}}{1 - 1/P_{1:N}} \tag{12a}$$

$$B = N + 1 - A - (N - A)/P_{N:N} \tag{12b}$$

The solution to Eq. (12) fits a straight line through the top and bottom ranks, assuming that all the intermediate ranks fall onto this model line, i.e., that Eq. (11) is linear with respect to $\langle y_{m:N}\rangle$ for all $m$. This paper refers to the solution of Eq. (12) as the "Gringorten extended" (GEX) methodology.

In practical extreme value analysis, the estimator is applied to all the ranks and a fit is made to all the ranked values – typically by a weighted least-mean-square (WLS) fit. It is therefore more appropriate to derive $A$ and $B$ by a fit to all ranks, when these are available, as this is consistent with their application in practice.

### 2.2 Three distributions useful in wind engineering

This paper addresses three distributions commonly used in wind engineering applications

1) FT1 distribution: $P = \exp\left(-\exp\left(-\frac{x-U}{C}\right)\right)$

2) Exponential distribution: $P = 1 - \exp(x/\bar{x})$

3) Weibull distribution: $P = 1 - \exp\left(-\left(\frac{x}{C}\right)^{w}\right)$

The FT1 distribution is used as the model for extreme events, typically for small to moderate $N$, and is the distribution addressed by Gringorten (1963). The Exponential distribution is the asymptote for the upper tail of the FT1 distribution and is the model for the time between events in a process following the Poisson recurrence model, again typically for small to moderate $N$. The Weibull distribution is commonly used as the model for the distribution of parent wind speeds, typically for very large $N$. These three distributions are so closely related that there is a strong commonality in the corresponding coefficient $A$ and $B$ values in Eq. (11).

### 2.3 The Fisher Tippett Type 1 distribution

The function $y = -\ln(-\ln(P))$ is the plotting position for the linearised expression for the FT1 distribution, Eq. (9). For the general case for rank $m$, Eq. (2) cannot be evaluated analytically, so requires a numerical approach, e.g., numerical integration as used by Harris (1999) or Bootstrapping. However, there are analytical expressions for the top and bottom ranks, enabling the GEX methodology to be applied and these are given in Appendix A.

Eq. (A.4) demonstrates that the coefficient $A$ has no significant effect on the top rank, $\langle y_{N:N}\rangle$, so that the upper tail of the distribution depends on the value of $B$. The coefficient $A$ has the greatest influence when $m$ is small, so controls the lower tail of the distribution. Eq. (12(a)) uses $P_{1:N}$ to solve for $A$ and Eq. (12(b)) uses $P_{N:N}$ to solve for $B$, giving the best conditioning of the simultaneous equations. A solution is quickly found by starting with $A = 0$ in Eq. (12(b)) to obtain a first estimate of $B$, then iterating between the two expressions[†]. Appendix A shows that $A \rightarrow B \rightarrow$

[†] Easily implemented in Microsoft Excel by using "circular references".

$1{-}e^{-\gamma} = 0.43854$ as $N \rightarrow \infty$. Gringorten's assumption that $A = B$, adopted from earlier work, implies that $N$ is large enough for both tails to be fully converged to the FT1 asymptote.

The analytical expressions for $\langle y_{1:N}\rangle$ and $\langle y_{N:N}\rangle$ are given in Gumbel (1958) but, while the expression for $\langle y_{N:N}\rangle$ on p210 is exact, the expression for $\langle y_{1:N}\rangle$ on p205 for the bottom rank is asymptotic and accurate to 2 decimal places only for $N > 10^6$. An improved expression for $\langle y_{1:N}\rangle$, which is accurate to 2 decimal places for $N > 20$ and to 3 decimal places for $N > 2000$, is given in Appendix B. Bootstrapped values of $\langle y_{1:N}\rangle$ were used to extend the range down to $N = 10$.

The resulting coefficients $A$ and $B$ for the FT1 distribution are shown converging to the asymptotic value in Fig. 3. All values were evaluated by GEX, except for the values labelled "$B$ – Bootstrap (WLS)" which were obtained from a WLS fit to all ranks.

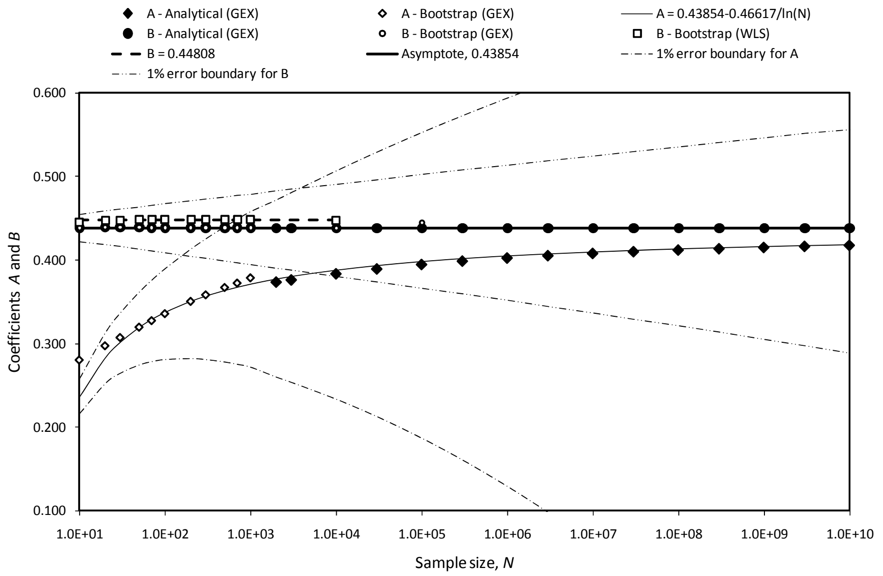


Fig. 3 The coefficients $A$ and $B$ in Eq. (11) for the FT1 distribution, converging to the asymptotic values with increasing sample size, $N$

### 2.3.1 Coefficient A.

Values of the coefficient $A$ are shown by the diamond symbols – solid symbols for the analytical expressions ($N \geq 2000$) and open symbols for the Bootstrapped values ($N \leq 1000$) from $10^8$ trials. These converge very slowly towards the asymptotic value, with error $O(1/\ln(N))$, and are still not close by $N = 10^{10}$. This behaviour is well fitted by $A = 0.43854 - 0.46617/\ln(N)$ which is shown by the solid-line curve. The two "dash-dot" curves above and below this fit represent the range of $A$ over which the error in $\langle y_{1:N}\rangle$ is less than 1% and indicate how the estimator becomes rapidly insensitive to the value of $A$ as $N$ increases. This slow convergence and insensitivity are due to the first double-logarithm term in Eq. (A.1), Appendix A, which dominates in the lower tail.

### 2.3.2 Coefficient B.

Values of the coefficient $B$ are shown by the circle symbols – again, solid for analytical and open for Bootstrapped values from $10^8$ trials. As the analytical values of $\langle y_{N:N}\rangle$ are exact and insensitive to $A$, the corresponding values of $B$ extend down to $N$ = 10. In contrast with $A$, the convergence of $B$ is extremely rapid indeed. The "dash-dot-dot" lines that diverge linearly above and below the asymptotic value represent the range of $B$ over which the error in $\langle y_{N:N}\rangle$ is less than 1% and again indicate how the estimator becomes less sensitive to the value of $B$ as $N$ increases.

The open square symbols show the value of $B$ obtained from a WLS fit to the Bootstrapped values from $10^8$ trials for all ranks. These values are consistently 2% larger than those from the GEX methodology operating on the same data.

### *2.4 The exponential distribution*

The function that linearises the Exponential distribution is

$$\langle y\rangle = \langle -\ln(1-P(x))\rangle = \frac{\langle x\rangle}{\bar{x}} \tag{13}$$

The upper tail is asymptotic to the FT1 and the lower tail is limited at $y = 0$.

Gumbel (1958, p117) gives the exact expression for the mean variate of the Exponential distribution for any rank, $m$, as

$$\langle y_{m:N}\rangle = \sum_{t=N+1-m}^{N} \frac{1}{t} \tag{14}$$

Eq. (14) should be implemented in the highest floating-point precision available to avoid accumulating rounding errors when $N$ is very large. Alternatively, use the expression of Harris (2009, Eq. 4.15), employing the digamma function. This exact expression implies there is no actual need for an estimator for the Exponential distribution. Note that Eq. (14) reverts to a single term for the lowest rank, $\langle y_{1:N}\rangle = 1/N$. Proofs are provided in Appendix A showing that $A$ may be defined as $A \equiv 0$ for all $N$, and that $B \to 0.43854$ as $N \to \infty$.

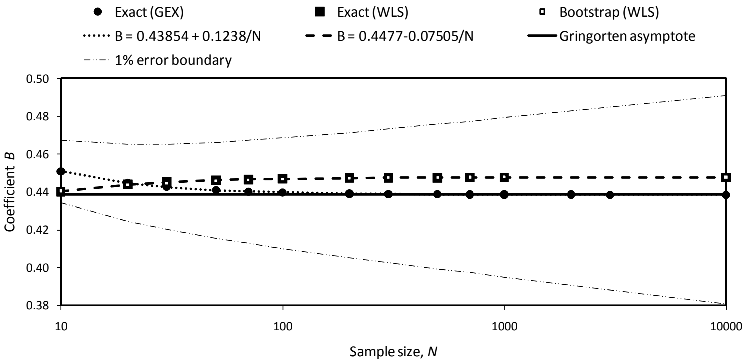


Fig. 4 The coefficient $B$ in Eq. (11) for the Exponential distribution, converging to the asymptotic values with increasing sample size, $N$

The resulting values of *B* for the Exponential distribution are shown in Fig. 4. The circle symbols denote values evaluated by GEX from the exact values for the top and bottom ranks. The square symbols denote values obtained from a WLS fit to all ranks: solid symbols from the exact expression, Eq. (14), and the open symbols from the Bootstrapped values from $10^8$ trials. Note that the Bootstrapped values and the exact analytic values are almost coincident.

Convergence of the values to the asymptote is slower than for the FT1 distribution. The WLS fits to all ranks are asymptotic to a value ~2% greater than the expected asymptote. Values from GEX converge from above, while the WLS values converge from below. The two "dash-dot-dot" curves above and below the GEX values represent the range of *B* over which the error in $\langle y_{N:N}\rangle$ is less than 1% and show that the difference in *B* between GEX and WLS has a very small effect on the values of $\langle y_{m:N}\rangle$.

### *2.5 The Weibull distribution*

$$P = 1 - \exp\left(-\left(\frac{x}{C}\right)^{w}\right) \tag{15}$$

where w is the shape parameter and C is the scale parameter, is one of the most ubiquitous statistical distributions used to describe physical phenomena. In wind engineering it is conventionally used to model parent hourly wind speed data for both hourly mean and hourly maximum gust values. The population, N, is typically very large and the standard WMO practice for rounding wind speeds to integer knot values leads to many tied values at low to moderate wind speeds. It is typical practice to reduce the size of the data set used in the analysis by using only the median rank, $\widetilde{m}(V)$, for each wind speed value, working from a frequency table – i.e., counts of each wind speed value – rather than the whole ranked set of observations, and this has implications for the appropriate fitting weights.

The function that linearises the Weibull distribution is

$$\langle y\rangle = \langle \ln(-\ln(1-P(x)))\rangle = w\langle \ln(x)\rangle - w\ln(C) \tag{16}$$

and differs from Eq. (9) and (13) in that $\langle y\rangle$ is the non-dimensional form of $\langle \ln(x)\rangle$. This function is diagonally symmetrical to the FT1 distribution, $y = -\ln(-\ln(P))$, such that

$$\begin{aligned}\langle y_{1:N}\rangle_{Weibull} &= -\langle y_{N:N}\rangle_{FT1}\\ \langle y_{m:N}\rangle_{Weibull} &= -\langle y_{N+1-m:N}\rangle_{FT1}\\ \langle y_{N:N}\rangle_{Weibull} &= -\langle y_{1:N}\rangle_{FT1}\end{aligned} \tag{17}$$

i.e., the signs are reversed and the rank is reversed top–bottom, equivalent to a diagonal reflection about a line of slope –1 through the origin. Hence values of $\langle y_{m:N}\rangle$ for the Weibull distribution at *m* = *i* may be obtained by evaluating $\langle y_{m:N}\rangle$ for the FT1 distribution at *m* = *N*+1–*i*, then reversing the sign. Alternatively, this diagonal reflection is achieved in the CLUE, Eq. (11), by exchanging the

FT1 values of the coefficients $A$ and $B$: if $A = a$ and $B = b$ for FT1, then $A = b$ and $B = a$ for the Weibull distribution.

### *2.6 Coefficient values for implementing CLUE*

Values of the Coefficients $A$ and $B$ for implementing Eq. (11) for the three distributions considered in this paper are presented in Table 1. These are based on the WLS fits, instead of the GEX methodology, because this better represents their use in practice. The two methodologies differ because the CLUE is not perfectly linear and the WLS fit is weighted to the middle of the distribution, while GEX uses only the endpoints. However, in practical applications, this difference is insignificant.

Table 1 Coefficients in Eq. (11) for implementing the CLUE for the three distributions

| Distribution | | Exponential | FT1 | Weibull |
|---|---|---|---|---|
| | | $\langle y\rangle = \langle -\ln(1-P)\rangle$ | $\langle y\rangle = \langle -\ln(-\ln(P))\rangle$ | $\langle y\rangle = \langle \ln(-\ln(1-P))\rangle$ |
| Eq. (11) | $A$ | 0 | 0.439–0.466/ln($N$) | 0.448 |
| | $B$ | 0.448–0.0751/$N$ | 0.448 | 0.439–0.466/ln($N$) |

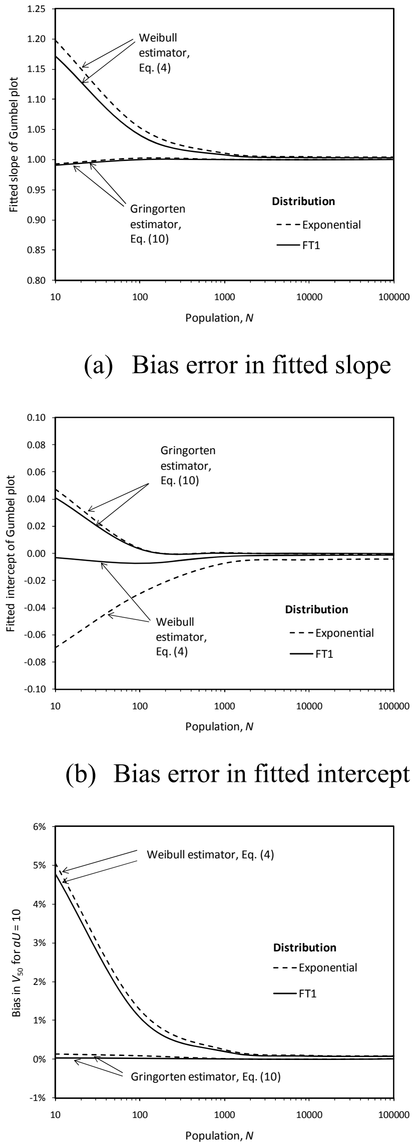
### *2.7 Calibration of the Weibull and Gringorten estimator bias errors*

The Weibull estimator for $\langle P\rangle$ is systematically biased in terms of $\langle x\rangle$ for all practical distributions $P(x)$, as discussed in *1.4*. The Gringorten estimator seeks to remove this bias for the FT1 distribution using the asymptotic values of the coefficients $A = B = 0.44$. The consistent linear unbiased estimator (CLUE) derived in this paper applies, with appropriate coefficient values, to FT1, Exponential and Weibull distributions and is accurate to better than 1% of $\langle x\rangle$ for any population of events, $N$.

Fig. 5 shows the bias in (a) the slope (cf. unity) and (b) the intercept (cf. zero) of the reduced variate for the Gringorten and Weibull estimators for the FT1 and Exponential distributions. For the Weibull distribution the diagonal symmetry with the FT1 distribution means that the bias in the slope is the same value and the error in the intercept is the negative value. The corresponding slope and intercept for the CLUE, using the coefficients in Table 1, are not shown in Fig. 5 because they are indistinguishable from 1 and 0, respectively.

The bias in the slopes is very similar for each distribution. The bias in the intercepts is more variable. Most of the bias in the Weibull estimator comes from the slope, underestimating by 15% for $N = 10$ (which leads to an overestimation of design values), and this bias remains greater than 1% for populations less than $N = 1000$. Most of the bias in the Gringorten estimator is in the intercept, an error in $y$ of ~0.04 for $N = 10$. The corresponding standard deviation for the highest rank of $N = 10$ is 8.3%, so the fitting error using the Weibull estimator is twice the sampling error, whereas the fitting error for the Gringorten estimator is small compared with the sampling error.

The effect of these bias errors on predicted design values for the variate depends on the value adopted for the design risk and on the mode/dispersion ratio, *U/b*. For mean wind speeds in temperate climates the characteristic product *U/b* ≈ 10, and the percentage error in the “once in 50 year” wind speed, $V_{50}$, for this case is shown in Fig. 5 (c). For the populations of annual maximum wind speeds typically available for analysis, the bias error on $V_{50}$ from using the Weibull estimator falls from around +5% at $N$ = 10 to around +2% at $N$ = 40, but the bias error from the Gringorten estimator remains less than 0.2%. This is because the errors in the mode and dispersion tend to cancel out when using the Gringorten estimator.



(a) Bias error in fitted slope

(b) Bias error in fitted intercept

(c) Percentage error in $V_{50}$ for typical temperate wind climate (aU = 10)

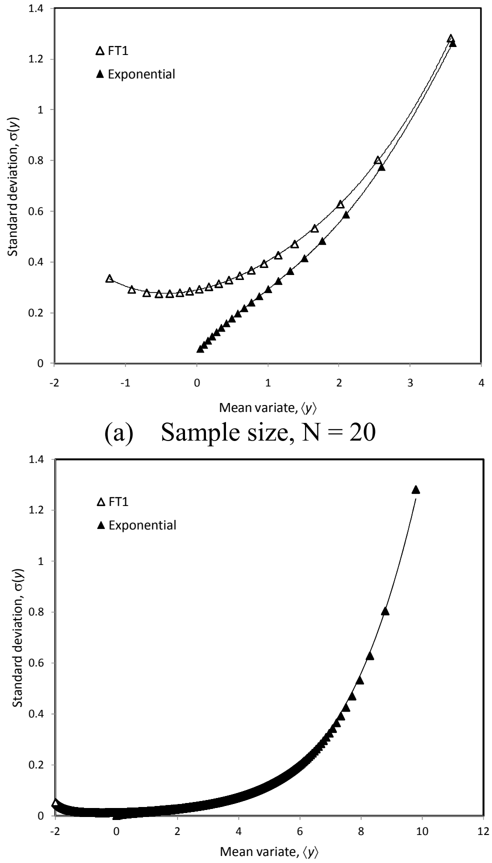
Fig. 5 Bias error in model fits using the Weibull and Gringorten estimators

## 3. Estimators for the standard deviations

For the estimator, Eq. (11), to be fully complemented with confidence limits and fitting weights, the standard deviation of the variate, $\sigma(\langle y_{m:N} \rangle)$, needs also to be estimated. For the Exponential distribution the standard deviations, $\sigma(y_{m:N})$, can be evaluated exactly (Gumbel 1958, p117) or via the derivative of the digamma function (Harris 2009, Eq. (5.4)). For the FT1 distribution, this is less easy to achieve analytically. Accordingly, for consistency of approach, Bootstrapping was used in this paper to derive the standard deviations for both distributions.

The standard deviations of the FT1 and Exponential reduced variates for $N = 20$ and $N = 10^4$ are shown in Fig. 6, from $10^6$ bootstrap trials, indicating that these are well fitted by 4th order polynomials

$$\sigma(y_{m:N}) = \prod\nolimits_{i=0}^{4} a_i \times (\langle y_{m:N} \rangle)^i \tag{18}$$



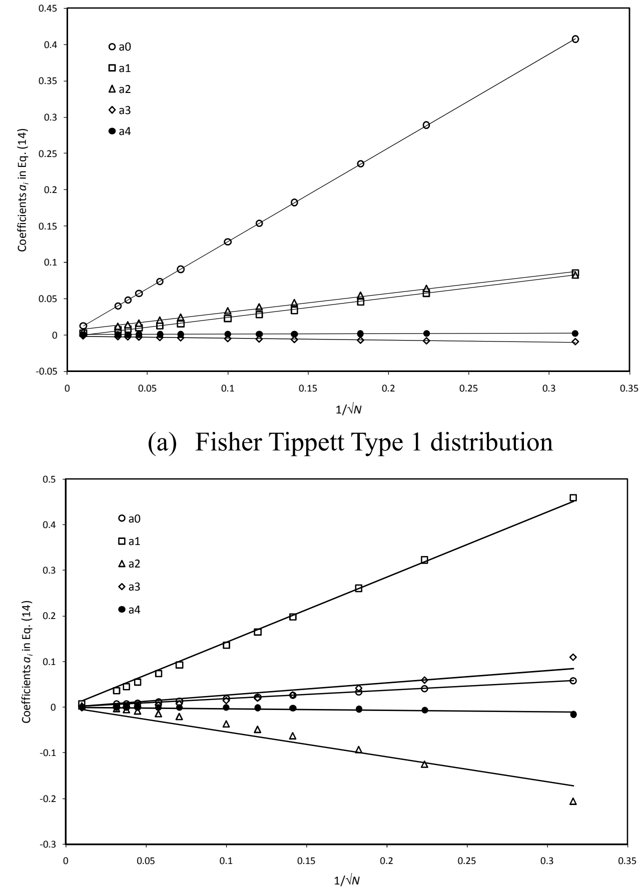
(a) Sample size, N = 20

(b) Sample size, N = $10^4$

Fig. 6 Standard deviations for $N = 20$ and $N = 10^4$ from $10^8$ bootstrap trials

The standard deviations for the Weibull reduced variate are the FT1 values, reflected around the $\langle y\rangle = 0$ axis, i.e. $\sigma(y)_{Weibull} = \sigma(-y)_{FT1}$. Fig. 7 shows that the coefficients $a_i$ are proportional to $1/\sqrt{N}$, leading to the expressions for each coefficient given in Table 2.

In a linear regression Castillo (1988) and Harris (1999) recommend that $y$ is taken as the dependent variable and $x$ as the independent variable, the reverse of the usual convention, because the principal uncertainty lies in the sampling error associated with $\langle y_{m:N}\rangle$. The standard deviations may be used to set confidence limits on the data and to provide fitting weights. To weight each value used in the fit commensurate with its statistical reliability, the square errors are multiplied by the weight corresponding to each rank. When using the full ranked data set, the weight is proportional to $1/\sigma^2$. When using only the median ranks for each wind speed, $\widetilde{m}(V)$, the weight is proportional to $|m(V)|/\sigma^2$, where $|m(V)|$ is the number of tied values of $V$. To evaluate the overall variance error of the fit, the weights should be normalised to sum to unity.



(a) Fisher Tippett Type 1 distribution

(b) Exponential distribution

Fig. 7 Coefficients $a_i$ in Eq. (14) from $10^8$ bootstrap trials

Table 2 Coefficients in Eq. (18) for the standard deviations for the three distributions

| Distribution | | Exponential<br>$\langle y\rangle = \langle -\ln(1-P)\rangle$ | FT1<br>$\langle y\rangle = \langle -\ln(-\ln(P))\rangle$ | Weibull<br>$\langle y\rangle = \langle \ln(-\ln(1-P))\rangle$ |
|---|---|---|---|---|
| Eq. (14) | $a_0$ | 0.202/√*N* | 1.279/√*N* | 1.279/√*N* |
| | $a_1$ | 1.284/√*N* | 0.235/√*N* | –0.235/√*N* |
| | $a_2$ | –0.294/√*N* | 0.334/√*N* | 0.334/√*N* |
| | $a_3$ | 0.132/√*N* | –0.0541/√*N* | 0.0541/√*N* |
| | $a_4$ | –0.0081/√*N* | 0.0121/√*N* | 0. 0121/√*N* |

## 4. Conclusions

This paper has explained the derivation of the Gringorten estimator and has described an improved estimator that is applicable to a range of distributions commonly used in wind engineering.

For the two distributions used in extreme value analysis, FT1 and Exponential, the classical Weibull estimator biases estimates extreme wind speeds around 4% too high for the typical range of observation periods available, while the original Gringorten estimator is accurate to better than 1%. We also show that the Gringorten estimator is also applicable to the Weibull distribution, which is typically used to represent parent winds where the population *N* is very high.

The CLUE in this paper, using the coefficients, *A* and *B*, provided in Table 1, is nearly two orders of magnitude more accurate than the Gringorten estimator. Accuracy for small sample sizes is increased significantly, but the sampling errors associated with the small sample sizes tend to swamp this improvement.

The coefficients, $a_i$, for the corresponding standard deviations, that have hitherto been unavailable, are useful for determining confidence limits and fitting weights.

## References

Castillo, E. (1988), *Extreme value theory in engineering*, Academic Press.

Cook, N.J. (2004), “Confidence limits for extreme wind speeds in mixed climates”, *J. Wind Eng. Ind. Aerod.*, **92**(1), 41-52.

Cramer, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press.

Efron, B. and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Gringorten, I.I. (1963), “A plotting rule for extreme probability paper”, *J. Geophys. Res.* **68**(3), 813-814.

Gumbel, E.J. (1958), *Statistics of Extremes*, Columbia University Press, New York, 371 pp.

Harris, R.I. (1999), “Improvements to the method of independent storms”, *J. Wind Eng. Ind. Aerod.*, **80**(1-2), 1-30.

Harris, R.I. (2009), “XIMIS – a penultimate extreme value method suitable for all types of wind climate”, *J. Wind Eng. Ind. Aerod.*, **97**(5-6), 271–286.

Naess, A. and Clausen, PH. (2001), “Combination of the peaks-over-threshold and bootstrapping methods for extreme value prediction”, *Struct.Saf.*, **23**(4), 315-330.

Weibull, W. (1939), “A statistical theory of the strength of materials”, *Proc. Roy. Swed. Acad. Eng. Sci.*, **151**, 5-45.

*JH*

## Appendix A: Asymptotic proofs for coefficients *A* and *B*

### *A.1 FT1 distribution*

### *A.1.1. Coefficient A*

Gumbel's asymptotic expression for the mean of the smallest ranked value from a FT1 distribution as $N \to \infty$ (Gumbel, 1958, p205) is

$$\langle y_{1:N} \rangle \approx -\ln(\ln(N)) - \gamma / \ln(N) \tag{A.1}$$

but is accurate only to the second decimal place for $N > 10^6$.

Substituting Eq. (11) into the FT1 linearised expression for $\langle y_{1:N} \rangle$, Eq.(9), gives

$$\begin{aligned} -\ln(\ln(N)) - \gamma / \ln(N) &= -\ln\left(-\ln\left(\frac{1-A}{N+1-A-B}\right)\right) \approx -\ln\big(\ln(N) - \ln(1-A)\big) \\ &\approx -\ln\left(\ln(N)\left(1 - \frac{\ln(1-A)}{\ln(N)}\right)\right) \approx -\ln(\ln(N)) + \frac{\ln(1-A)}{\ln(N)} + O\left(\frac{1}{\ln^2(N)}\right) \end{aligned} \tag{A.2}$$

Hence: $\ln(1-A) = -\gamma$ as $N \to \infty$

Thus: $A = 1 - e^{-\gamma} = 0.43584$ as $N \to \infty$

### *A.1.2. Coefficient B*

Given the exact value of $\langle y_{N:N} \rangle = \gamma + \ln(N)$ for a FT 1 distribution – see (Gumbel 1958, p210), where $\gamma = 0.5772157...$ is Euler's constant

$$\gamma + \ln(N) = -\ln\left(-\ln\left(\frac{N-A}{N+1-A-B}\right)\right) = -\ln\left(-\ln\left(1 - \frac{A}{N}\right) + \ln\left(1 + \frac{1}{N} - \frac{A}{N} - \frac{B}{N}\right)\right) \tag{A.3}$$

Expanding the inner ln() terms as series gives

$$\begin{aligned} \gamma + \ln(N) &= -\ln\left(\frac{A}{N} + \frac{1}{N} - \frac{A}{N} - \frac{B}{N} + O\left(\frac{1}{N^2}\right)\right) \\ &= -\ln\left(\frac{1-B}{N} + O\left(\frac{1}{N^2}\right)\right) \approx +\ln(N) - \ln(1-B) \end{aligned} \tag{A.4}$$

Hence: $B \to 1 - e^{-\gamma} = 0.43854...$ as $N \to \infty$.

### *A.2. Exponential distribution*

### *A.2.1. Coefficient A*

Substituting the exact expression $\langle y_{1:N} \rangle = 1/N$ and Eq. (11) into Eq. (13) gives

$$\begin{aligned}\frac{1}{N} &= -\ln\left(1-\frac{1-A}{N+1-A-B}\right) = -\ln\left(\frac{N-B}{N+1-A-B}\right) \\ &= -\ln(1-B/N)+\ln(1+1/N-A/N-B/N)\end{aligned} \tag{A.5}$$

Expanding both the ln() terms as series gives

$$\begin{aligned}&\frac{1}{N} = \frac{B}{N}+\frac{1}{N}-\frac{A}{N}-\frac{B}{N}+O\left(\frac{1}{N^2}\right) \\ &\approx \frac{1}{N}-\frac{A}{N}\end{aligned} \tag{A.6}$$

Hence $A/N \cong 0$ for any $N$ where $O(1/N^2)$ is negligible, meaning that $A << N$. Hence one can always take $A \cong 0$ and accommodate any residual second-order effects into the value of $B$.

### *A.2.2. Coefficient B*

Given that the asymptotic value of $\langle y_{N:N}\rangle = \gamma + \ln(N)$ – see (Gumbel 1958, p116), where $\gamma =$ 0.5772157... is Euler's constant

$$\begin{aligned}\gamma+\ln(N) &\approx -\ln\left(1-\frac{N-A}{N+1-A-B}\right) = -\ln\left(\frac{1-B}{N+1-A-B}\right) \\ &= -\ln(1-B)+\ln(N)+\ln(1+1/N-A/N-B/N)\end{aligned} \tag{A.7}$$

Expanding the last ln() term as a series gives

$$\gamma \approx -\ln(1-B)-\frac{1}{N}+\frac{A}{N}+\frac{B}{N}+O\left(\frac{1}{N^2}\right) \approx -\ln(1-B) \tag{A.8}$$

Hence: $B \to 1 - e^{-\gamma} = 0.43854...$ as $N \to \infty$.

## Appendix B: Improved asymptotic expression for the mean FT1 variate of the smallest rank

An improved asymptotic expression for the mean of the smallest ranked value from a FT1 distribution is derived by expanding Eq. (2) as a series and integrating term by term, as follows.

From Eq. (2)

$$\langle y_{1:N} \rangle = N \int_0^1 -\ln(-\ln z)(1-z)^{N-1} dz \tag{B.1}$$

Let $z = t/(N–1)$ and $dz = dt/(N–1)$, then

$$\begin{aligned}
\langle y_{1:N} \rangle &= \frac{N}{N-1} \int_0^{N-1} -\ln(-\ln(N-1) - \ln t)\left(1 - \frac{t}{N-1}\right)^{N-1} dt \\
\text{using Cauchy ...} \quad &\approx \frac{N}{N-1} \int_0^{N-1} -\ln\left(-\ln(N-1)\left(1 - \frac{\ln t}{\ln(N-1)}\right)\right) e^{-t} dt \\
&\approx \frac{N}{N-1} \int_0^{\infty} \left(-\ln(-\ln(N-1) + \frac{\ln t}{\ln(N-1)} + \frac{1}{2}\left(\frac{\ln t}{\ln(N-1)}\right)^2 \ldots\right) e^{-t} dt
\end{aligned} \tag{B.2}$$

Now integrating term by term using standard solutions in Gradshteyn and Ryzhik (1969)

$$\begin{aligned}
\langle y_{1:N} \rangle &\approx \frac{N}{N-1}\left[-\ln(-\ln(N-1)) - \frac{\gamma}{\ln(N-1)} + \frac{\pi^2/6 + \gamma^2}{2\ln^2(N-1)} \ldots\right] \\
&\approx \frac{N}{N-1}\left[-\ln(-\ln(N-1)) - \frac{0.5772157}{\ln(N-1)} + \frac{0.9890560}{\ln^2(N-1)} \ldots\right]
\end{aligned} \tag{B.3}$$

This expression has one more term than the Gumbel expression and is accurate to the second decimal place for $N > 20$ and to the third decimal place for $N > 2000$.

*Reference:* Gradshteyn, I.S., Ryzhik, I.M., 1969, Tables of Integrals, Series and Products, Academic Press, New York.