Example of an approximately normally distributed data set to which a large number of different probability distributions can be fitted

R.J. Oosterbaan 15-10-2019. On www.waterlog.info public domain.

Abstract

A data set is shown that obeys a standard normal probability distribution but to which ten other probability distributions have been fitted that all reveal high R^2 values (a goodness of fit parameter) of 0.983 to 0.987

Contents

1. Normal distribution, standard, $R^2 = 0.985$

The CumFreq software [Ref. 1] uses the cumulative distribution functions (CDF) of probability distributions and fits those to a data set using the Weibull (or Gumbel) plotting position [Ref. 2].
However, for the normal distribution the CDF is not known, but it can be approximated closely using the Hastings formula [Ref. 3]:

$$Fc = 1 - N(1\ 0.319\ Y - 0.357\ Y^2 + 1.781\ Y^3 - 1.821\ Y^4 + 1.330Y^5)$$

where

Fc = cumulative normal distribution function or cumulative normal frequency
$N = \{1/\sqrt{(2\pi)}\}\ e^{\{-(Z\wedge 2)/2\}}$
Z = 1/(1+0.232Y)
Y = (X – M) / StD
X = the random variable under study
M = the mean of the X values
StD = the standard deviation of the X values
e = base of the natural logarithm (Ln), e = 2.71 . . .
^ = symbol signifying: raised to the power 2
π = 3.141 . . . the ratio of the surface of a circle divided by the square value of its radius

Fc is normally found from the following equation (also called plotting position):

$Fc = R/(N+1)$,

where

R = the rank number of the respective X values arranged in an ascending order
N = the number of data

The result of fitting the cumulative normal distribution to the standard data set with 43 values used in this article is shown in the following figure.

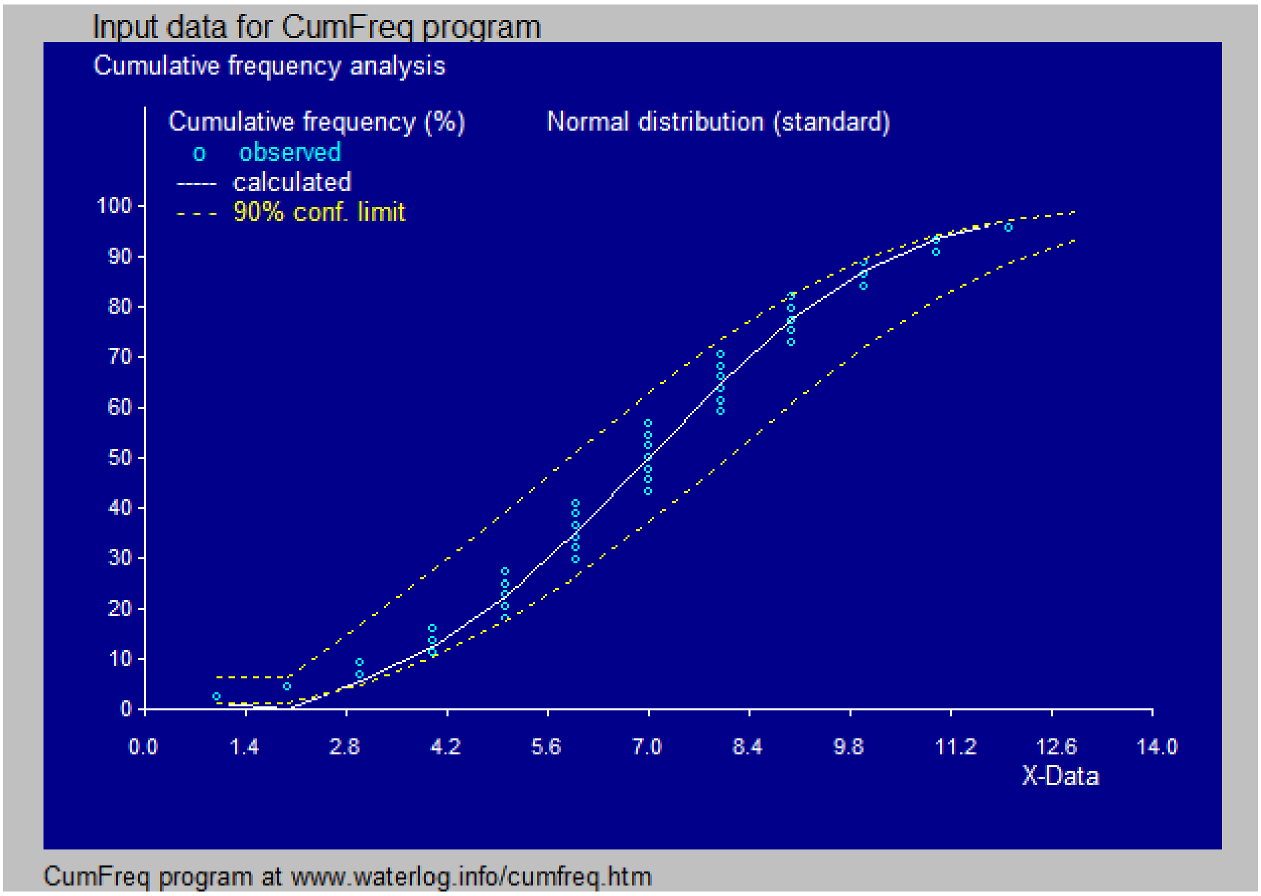

Figure 1. The standard normal distribution fitted to the standard data set used in this article.

The mean M of X equals 7.00 and the standard deviation StD of X equals 2.64.
The calculator accompanying the CumFreq program gives X = 13.14 for Fc = 0.99 or 99%

2. Generalized exponential distribution (Poisson type), $R^2 = 0.984$

The CDF of the generalized exponential distribution can be given as:

$Fc = 1 - \text{Exp}\{-(A*X\text{^}E+B)\}$

where E is an exponent that is to be found by optimization minimizing the sum of squares of deviations of calculated from predicted Fc values, or maximizing the R^2 value. The addition of exponent E to the standard exponential distribution is the basis of the generalization and adjusts the skewness of the distribution

Using the transformations:

$Xt = \text{Ln}(X\text{^}E) = E*\text{Ln}(X)$

$$Ft = -\mathrm{Ln}\,(1-Fc)$$

we find the linear relation:

$$Ft = A*Xt + B$$

where A and B can be found found from a linear regression of Ft upon Xt.

The result of fitting the generalized exponential distribution to the standard data set used in this article is shown in the following figure.

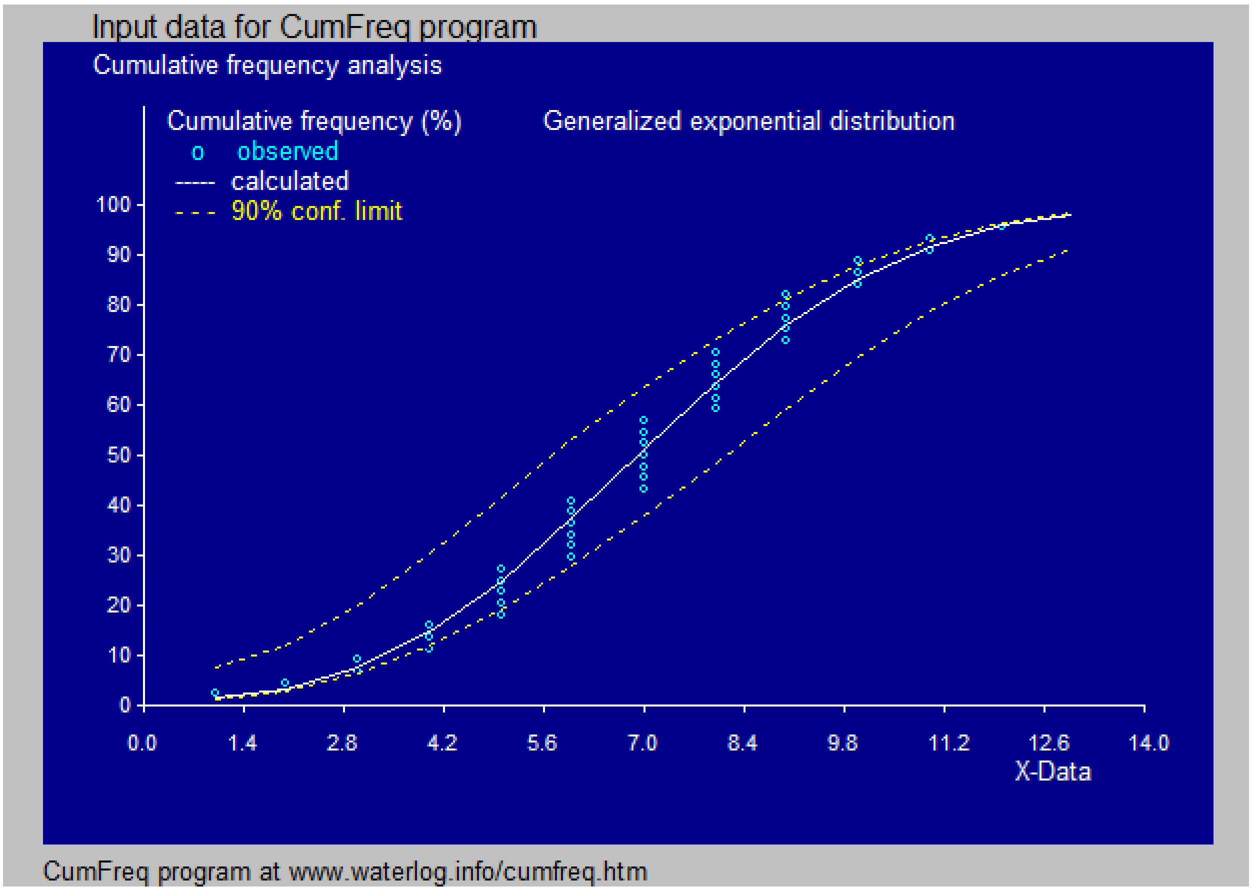

Figure 2. The generalized exponential distribution fitted to the standard data set used in this article.

The exponent E in this case = 2.79, while the parameters are A = 0.0031 and B = 0.0115
The calculator accompanying the CumFreq program gives X = 13.70 for Fc = 0.99 or 99%

3. Fisher-Tippett type 3 distribution, $R^2 = 0.984$

The CDF of the Fisher-Tippett type 3 distribution can be given as:

$$Fc = \mathrm{Exp}[-\{(C-X)/\mathrm{Exp}(-B/A)\}\hat{}A]$$

Using the transformations:

$$Xt = \mathrm{Ln}(C-X) \qquad [X<C]$$

$$Ft = \mathrm{Ln}\{-\mathrm{Ln}(Fc)\}$$

we find the linear relation:

$$Ft = A*Xt + B$$

where A and B can be found found from a linear regression of Ft upon Xt.
The value of parameter C (the maximum possible X) is to be optimized.

The result of fitting the generalized exponential distribution to the standard data set used in this article is shown in the following figure.

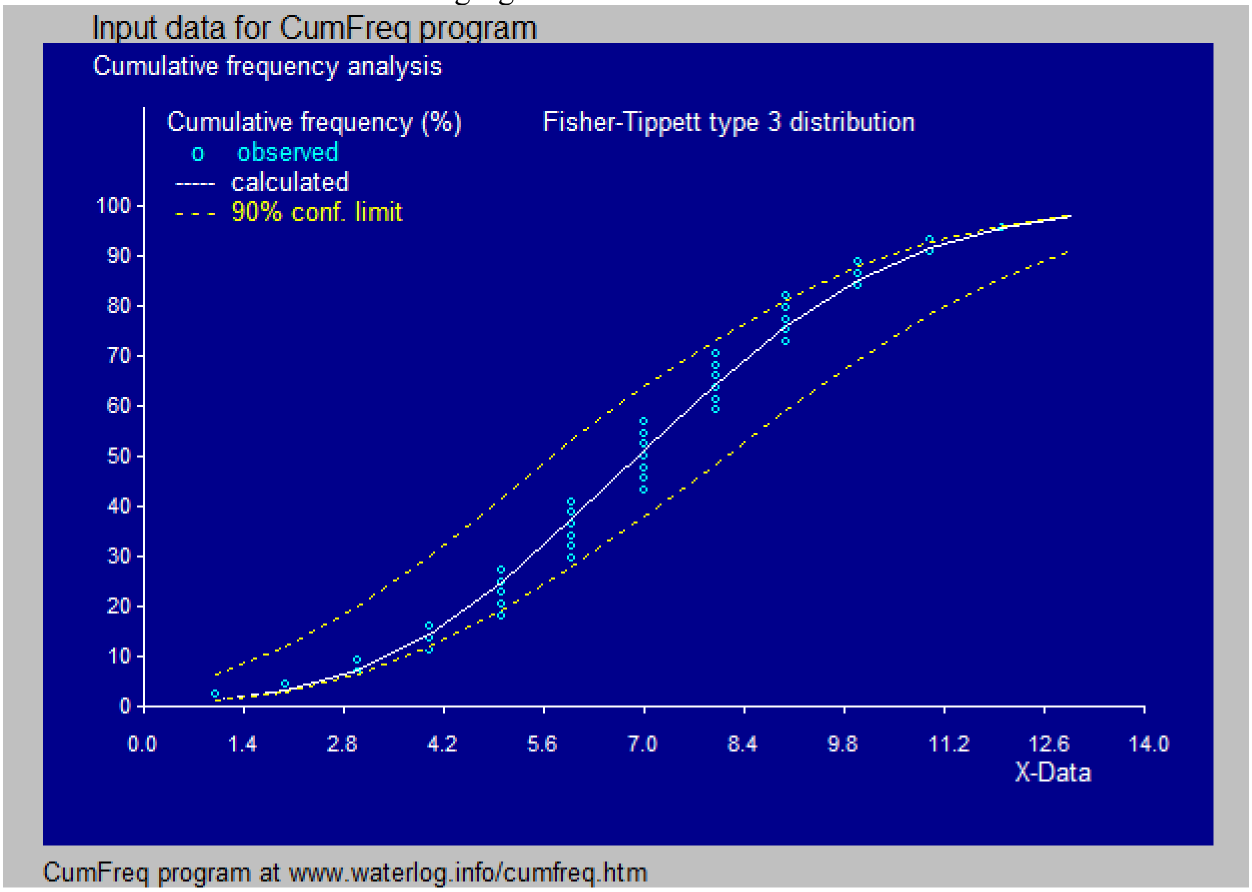

Figure 3. The Fisher-Tippett tye 3 distribution fitted to the standard data set used in this article.

The optimized value of C = 17.4, while the parameters are A = 4.202 and B = – 10.25
The calculator accompanying the CumFreq program gives X = 13.60 for Fc = 0.99 or 99%

4. Mirrored Frechet (Fisher-Tippett type 2) distribution, $R^2 = 0.985$

The CDF of the mirrored Frechet (Fisher-Tippett type 2) distribution can be given as:
Fc = 1–Exp[–{(X–C)/Exp(–B/A)}^A]
Using the transformations:
Xt = Ln(X–C) [X>C]
Ft = Ln{–Ln(Fc)}
we find the linear relation:
Ft = A*Xt + B
where A and B can be found found from a linear regression of Ft upon Xt.
The value of parameter C (the minimum possible X) is to be optimized.

The result of fitting the generalized exponential distribution to the standard data set used in this article is shown in the following figure.

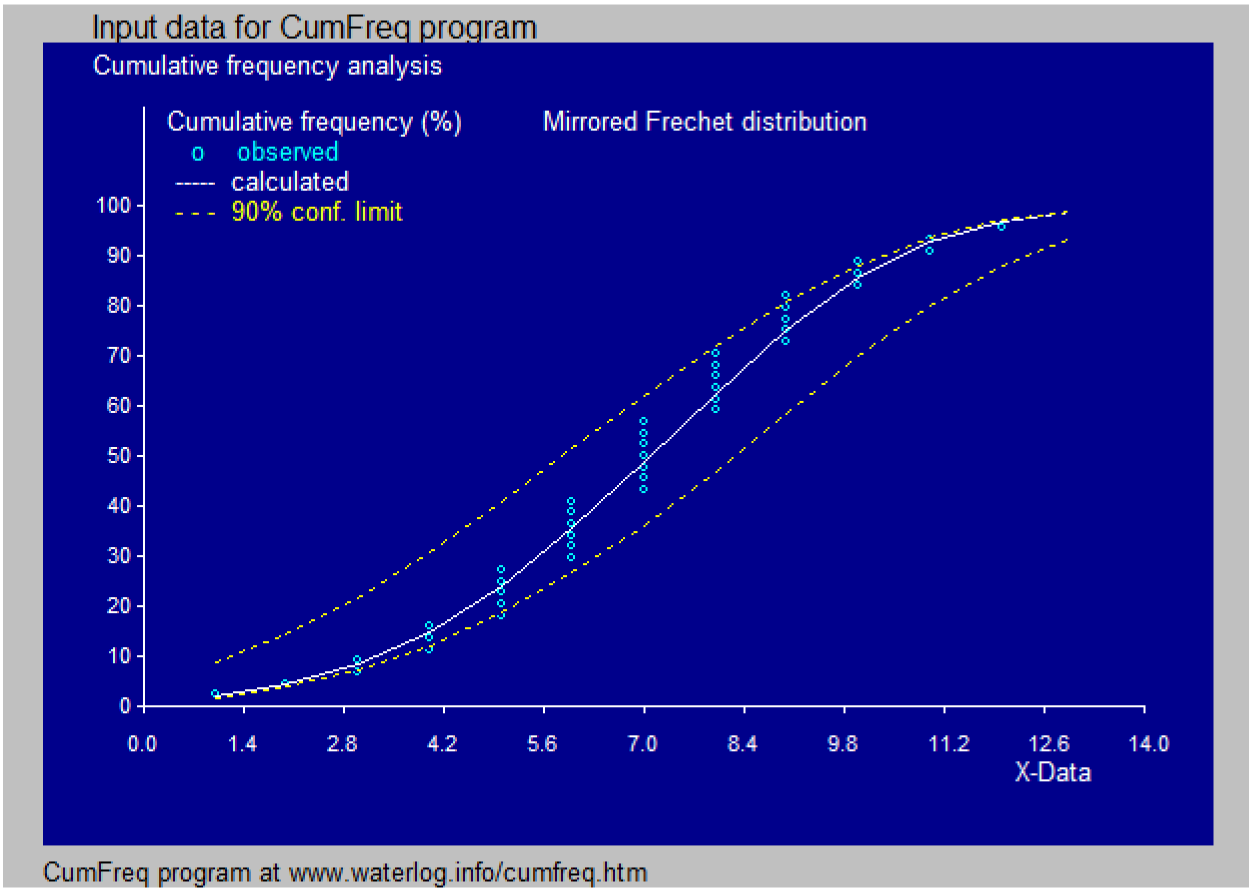

Figure 4. The mirrored Frechet distribution fitted to the standard data set used in this article.

The optimized value of C = – 3.44, while the parameters are A = 4.20 and B = – 10.25
The calculator accompanying the CumFreq program gives X = 13.60 for Fc = 0.99 or 99%

5. Kumaraswamy distribution, R^2 = 0.985

The CDF of the Kumaraswamy distribution can be given as:
Fc = 1 – {1 – (X/M)^A}^B
Using the transformations:
Xt = Ln{(X/M)^A}
= A*Ln(X/M) (use ratio method to find A *)
Ft = Ln(1–Fc)
we find the linear relation:
Ft = B*Xt (use ratio method to find B *)
The value of parameter M > Xmax is to be optimized
*) The ratio method is a linear regression by which the regression line is forced to go through the origin (where Xt=0 and Ft=0).

The result of fitting the generalized exponential distribution to the standard data set used in this article is shown in the following figure.

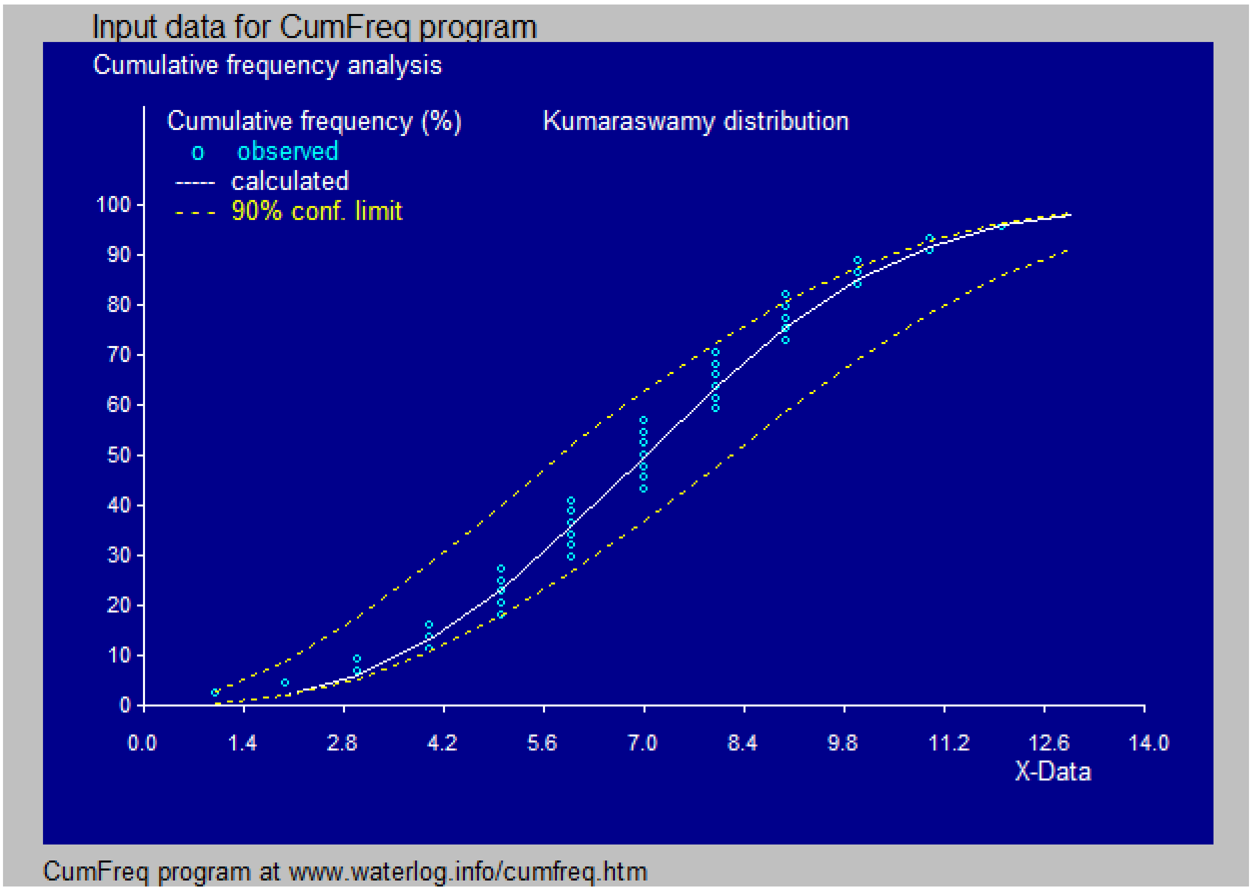

Figure 5. The Kumaraswamy distribution fitted to the standard data set used in this article.

The optimized value of M = 237, while the parameters are A = 2.84 and B = 22747

6. Generalized Burr distribution, R^2 = 0.985

The CDF of the generalized Burr distribution can be given as:
Fc = 1 – [{B/(X+B)}^A]^E [B>0, X>-B]
Using the transformations:
Xt = Ln[{B/(X+B)}^A]
Ft = Ln(1–Fc)
we find the linear relation:
Ft = E*Xt (use ratio method to find E #)
The values of A and B are to be optimized.
#) The ratio method is a linear regression by which the regression line is forced to go through the origin (where Xt=0 and Ft=0).

The result of fitting the generalized Burr distribution to the standard data set used in this article is shown in the following figure.

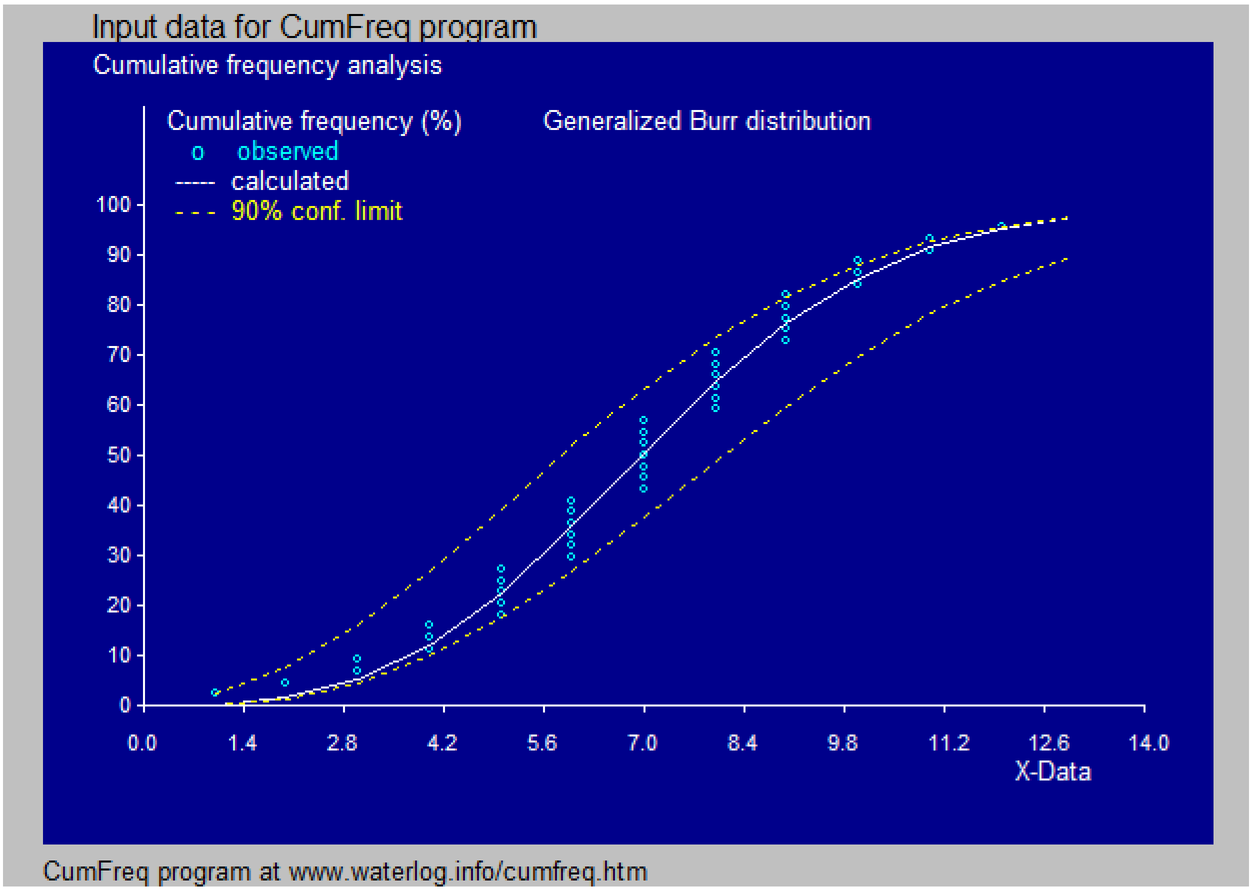

Figure t. The generalized Burr distribution fitted to the standard data set used in this article.

The value of exponent E is found with the ratio method as E = 5.30, while the values of A and B are optimized as A = 3.15 and B = 13.0

7. Generalized extreme value (GEV) distribution, R^2 = 0.986

The CDF of the GEV distribution (also called Fisher-Tippett type 1) can be given as:

$$Fc = \exp[-\{1+K(X-A)/B\}^{\wedge}(-1/K)]$$

where

K, A and B are parameters all to be optimized

The result of fitting the GEV distribution to the standard data set used in this article is shown in the following figure.

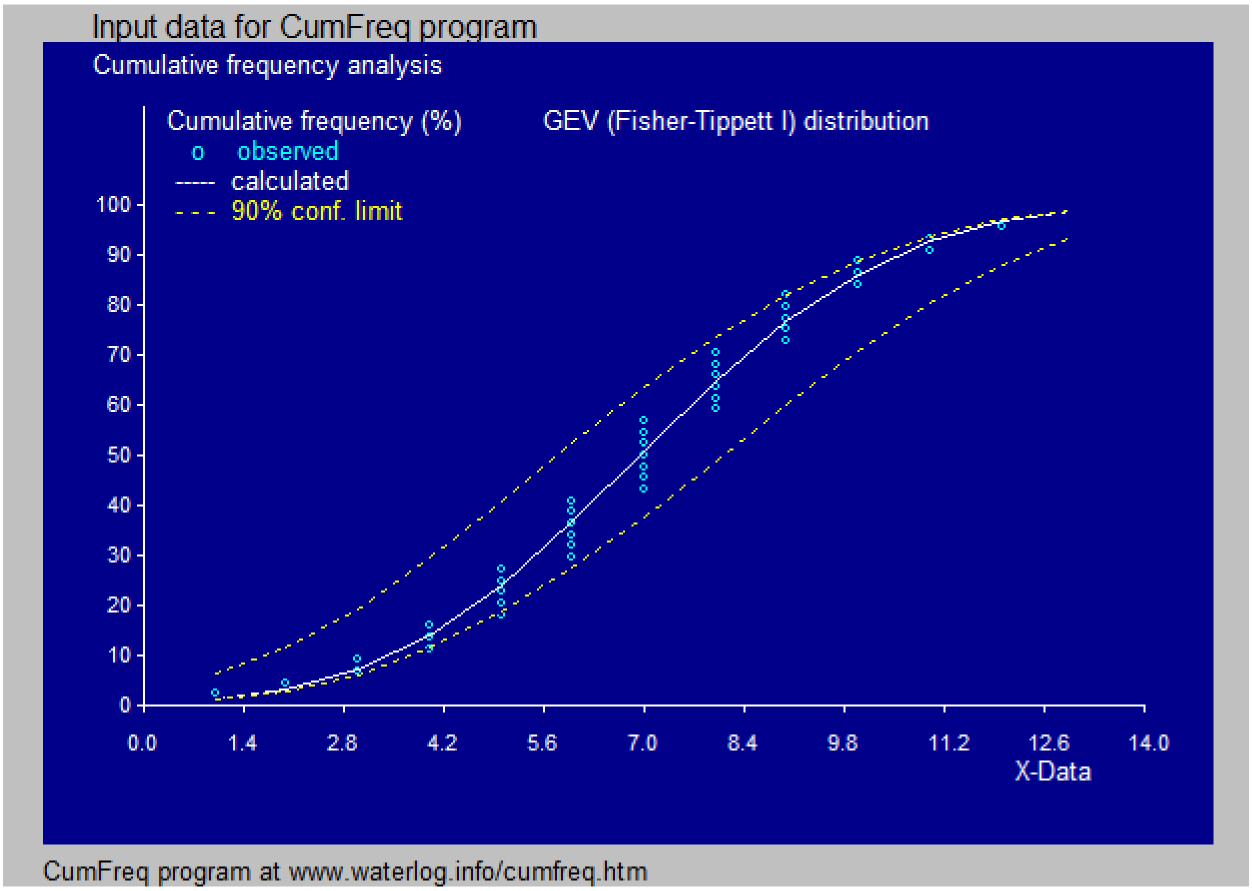

Figure 7. The GEV (Fisher-Tippett type 1) distribution fitted to the standard data set used in this article.

The optimized values of K, A and B are – 0.270, 6.00 and 2.69 respectively.

8. Normal distribution, optimized, $R^2 = 0.986$

The equations used are the same as presented in section 1, with the exception that the standard error of X (2.64) is optimized to 2.74

The result of fitting the optimized normal distribution to the standard data set used in this article is shown in the following figure.

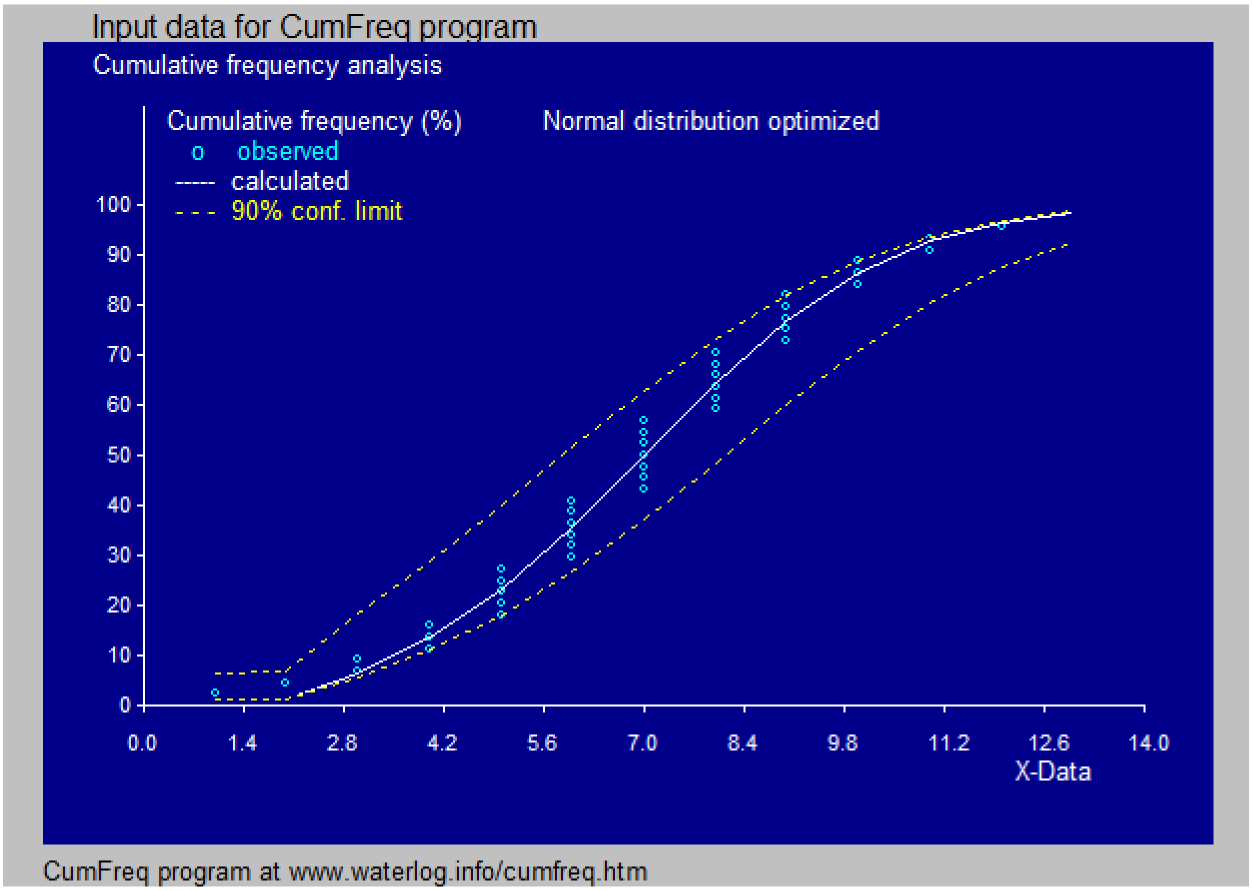

Figure 8. The optimized normal distribution fitted to the standard data set used in this article.

9. Generalized Gumbel distribution, R^2 =0.987

The CDF of the Gumbel distribution can be given as:
Fc = Exp[–Exp{– (A*X^E+B)}]

Using the transformations:
Xt = Ln(X^E) = E*Ln(X)
Ft = –Ln{–Ln(Fc)}
we find the linear relation:
Ft = A*Xt + B
where A and B can be found found from a linear regression of Ft upon Xt.
The value of exponent E is to be optimized.

The result of fitting the generalized Gumbel distribution to the standard data set used in this article is shown in the following figure.

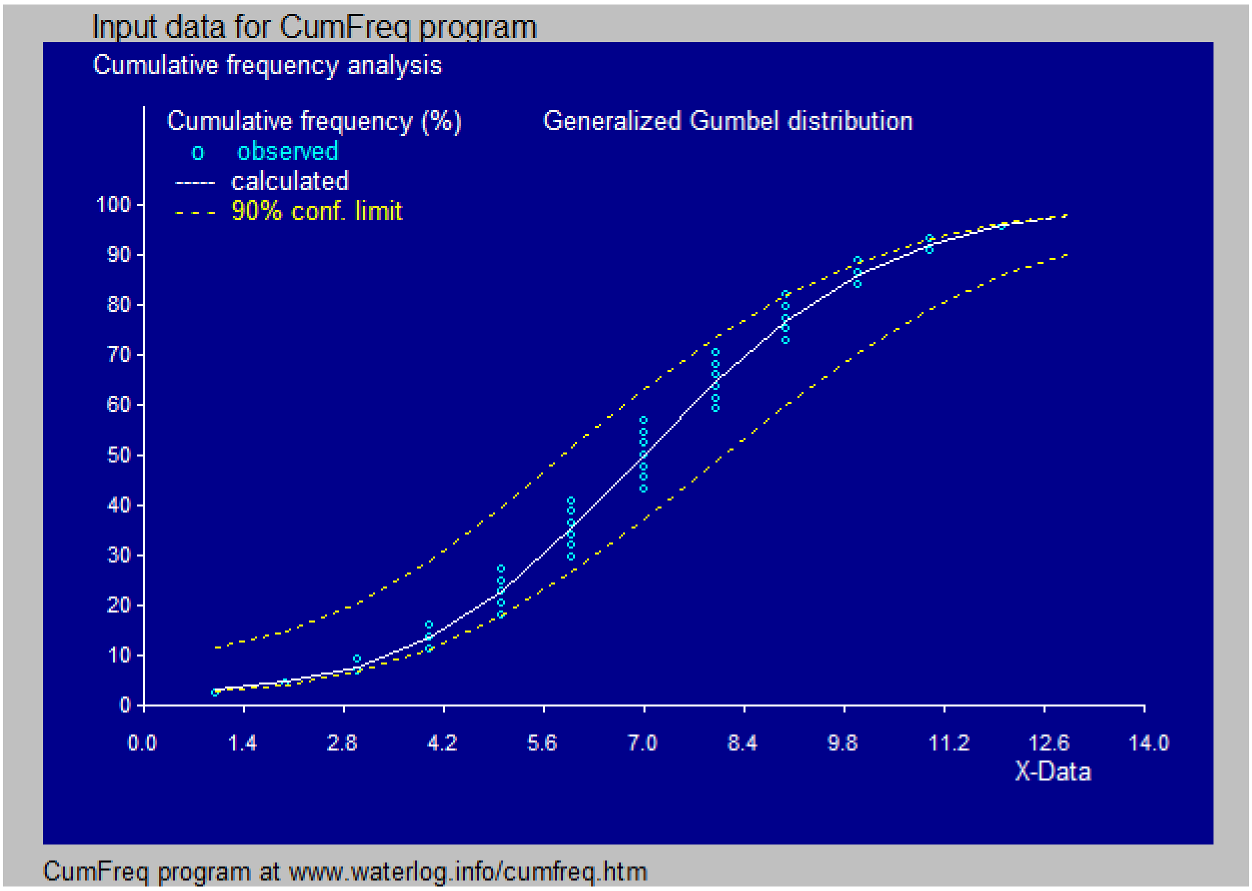

Figure 9. The generalized Gumbel distribution fitted to the standard data set used in this article.

The optimized value of E = 1.82, while the parameters are found by linear regression as A = 0.0843 and B = – 1.29
The calculator accompanying the CumFreq program gives X = 14.01 for Fc = 0.99 or 99%

10. Generalized Gumbel distribution mirrored, R^2 =0.983

The CDF of the mirrored Gumbel distribution can be given as:
Fc = 1-Exp[–Exp{– (A*X^E+B)}]

Using the transformations:
Xt = Ln(X^E) = E*Ln(X)
Ft = –Ln{–Ln(1+Fc)}
we find the linear relation:
Ft = A*Xt + B
where A and B can be found found from a linear regression of Ft upon Xt.
The value of exponent E is to be optimized.

The result of fitting the generalized Gumbel distribution to the standard data set used in this article is shown in the following figure.

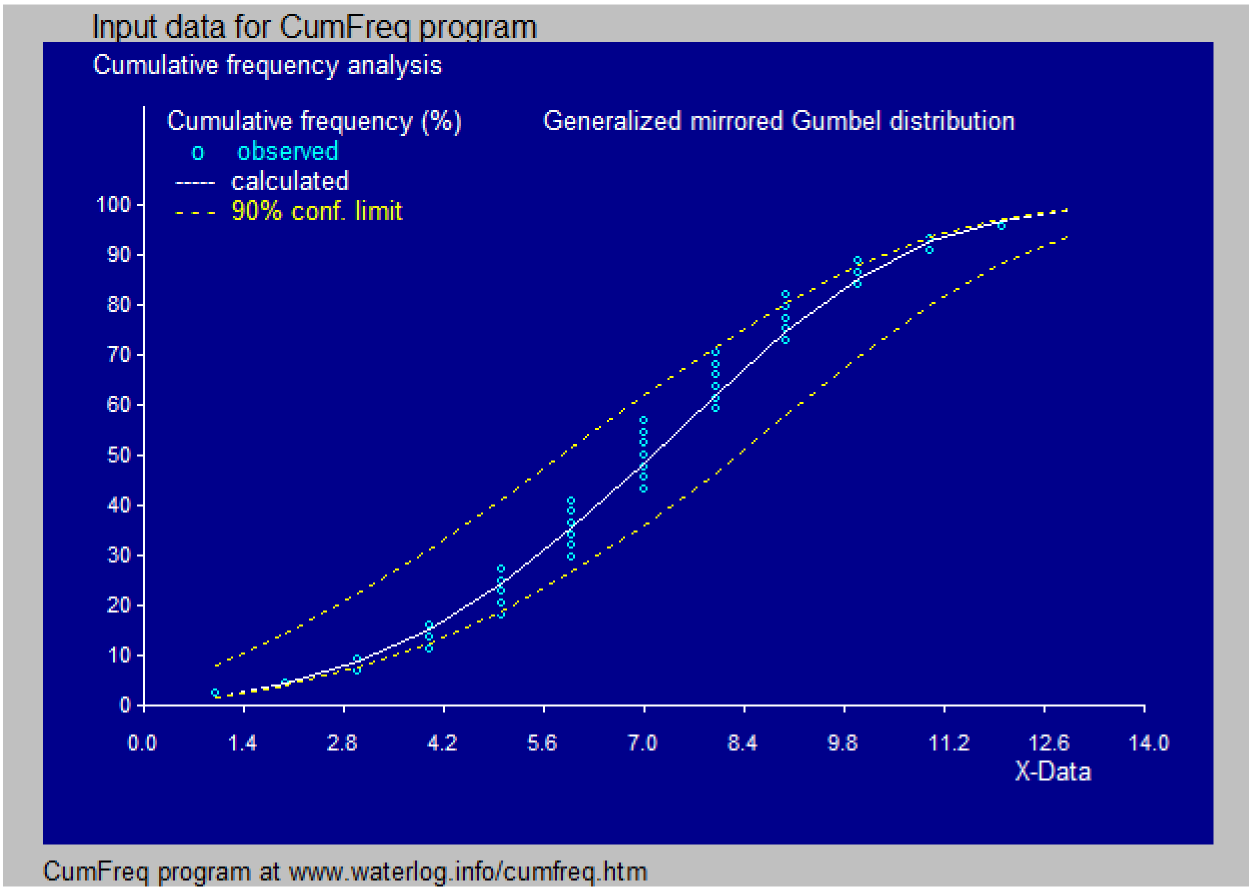

Figure 10. The generalized mirrored Gumbel distribution fitted to the standard data set used in this article.

The optimized value of E = 0.41, while the parameters are found by linear regression as A = – 3.03 and B = 7.14

11. Logistic distribution, R^2 = 0.987 (best of all)

The CDF of the logistic distribution can be given as:

Fc = 1/(1+Exp(A*X+B))

This CDF can be rewritten in linear form as:

Ln (1 / Fc) = A*X + B

so that the parameters A and B can be found from a linear regression of Y = Ln (1 / Fc) on X.

The versatility of the logistic distribution has been discussed before [Ref. 4]. The result of fitting the logistic distribution to the standard data set used in this article is shown in the following figure.

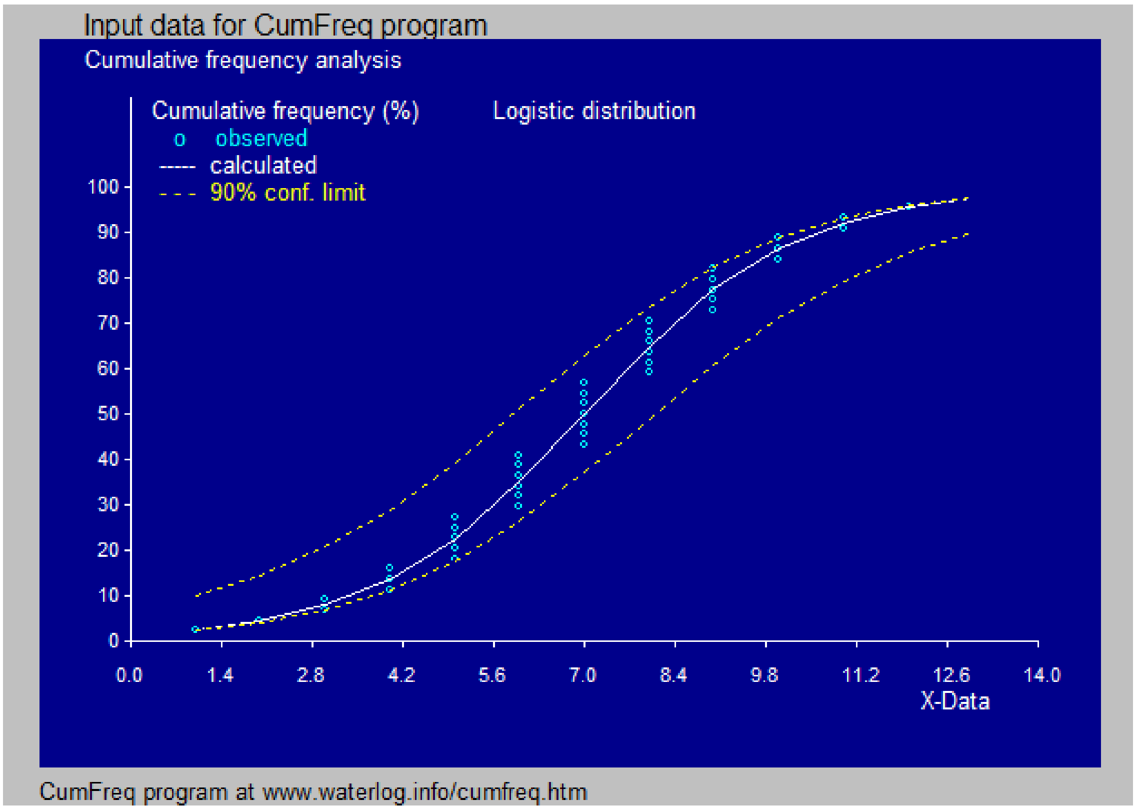

Figure 11. The standard logistic distribution fitted to the standard data set used here.

The values of the parameters found by linear regression are A = – 0.617576 and
B = 4.32303

11. Conclusion

The 11 graphs shown do not not manifest clear distinctions, they are all very similar and they all show an excellent fit.
Some of the characteristics of the distribution are summarized in table 1.

Table 1. Summary of some of the characteristics of the 10 probability distributions used in this article

Distribution	**R^2**	**X at Fc = 99% (extrapolation)**
1. Normal, standard	0.985	13.14
2. Generalized exponential	0.984	13.70
3. Generalized Fisher-Tippett type 3	0.984	13.60
4. Mirrored Frechet (mirrored Fisher-Tippett type 2)	0.985	13.01
5. Kumaraswamy	0.985	13.66
6. Generalized Burr	0.985	14.41
7. Generalized extreme value (GEV) (Fisher-Tippett type 1)	0.986	13.08
8. Normal distribution, optimized	0.986	13.38
9. Generalized Gumbel	0.987	14.01
10. Generalized Gumbel mrrored	0.983	12.97
11. Logistic distribution	0.987	14.44

Despite the large differences in the mathematical expressions of the PDF's, the correct determination of their parameters, either by linear regression after their linearization, or by optimization, i.e. maximizing the R^2, leads to fitting results that are quite similar and all have a high goodness of fit. Also the extrapolation of the PDF, gives similar values of around X = 13 or 14.

12. References

[Ref. 1]. CumFreq, free software for probability distribution fitting. Download from: https://www.waterlog.info/cumfreq.htm

[Ref. 2]. Definition of the Weibull plotting position. On line: http://glossary.ametsoc.org/wiki/Weibull_plotting_position

[Ref. 3]. M. Abramowitz, I.A. Stegun. Handbook of mathematical functions (10th ed.), Dover Publications, New York (1972), pp. 925-976

[Ref. 4]. Fitting the versatile linearized, composite and generalized logistic distribution. On line: https://www.researchgate.net/publication/335022301_FITTING_THE_VERSATILE_LINEARIZED_COMPOSITE_AND_GENERALIZED_LOGISTIC_PROBABILITY_DISTRIBUTION_TO_A_DATA_SET