The generalized standard and mirrored Gumbel probability distributions, composite or not, are applicable to many datasets, either symmetrical, skew to the left, or skew to the right

R.J Oosterbaan

June 2022 www.waterlog.info/cumfreq.htm

Abstract
The Laplace probability distribution is a well known composite distribution and mostly called a double distribution. The expression "double" is somewhat misleading because it concerns the division of the data set in two parts. Other composite probability distributions are not commonly dealt with in literature. However, using the Gumbel plotting position as cumulative frequency estimator, it is possible to fit data to any composite distribution with more or less success. Not only can the well known probability distributions as such be cut into two parts, but also the distribution in part 1 can be different from the one employed in the second part (mixed composition). The estimation of the distribution parameters can often be done by a transformation followed by a linear regression. The fit of the data to the (composite) distribution can be further enhanced by raising the data to a power that will have to be optimized, yielding a "generalized" probability distribution. Such a kind of transformation is well known from the log-normal distribution, but it can be employed for all other distributions. The concept of Mirrored probability distributions is also not frequently used in literature, but the Burr and Dagum probability distributions are an example as one is mirrored with respect to the other. This paper will give composite probabilities using the generalized standard and mirrored Gumbel distribution,. It will also give examples situations in which composition is advisable. The overview and the examples will be presented with the help of the free CumFreqA model software, made for the purpose of generalization and composition, as calculations by hand would be cumbersome.

Contents

1. Introduction

1.1 Methods of distribution fitting

The following techniques of distribution fitting exist:

- *Parametric methods*, The parametric methods are:
  - method of moments
  - maximum spacing estimation
  - method of L-moments
  - Maximum likelihood method

- *Regression method*, using a transformation of the cumulative distribution function so that a linear relation is found between the cumulative probability and the values of the data, which may also need to be transformed, depending on the selected probability distribution. In this method the cumulative probability needs to be estimated by the plotting position.

In case the regression method is not applicable, the *numerical method* of parameter optimization can be used instead of the parametric methods.

1.2. The Gumbel plotting position

The Gumbel plotting position (Pp) gives an estimate of the cumulative probability (Cp or probability of non-exceedance) for each of the values in a data set. Before the Pp can be determined the data set must be arranged in ascending order. Each value Xn in this series with n = 1, 2, 3, ... N (where N is the total number of data) is given the PP value n / (N+1).

Gumbel (1954, *Ref. 1*) has shown that Pp is an unbiased estimator of the cumulative probability around the mode of the distribution. In literature there exist other estimates, but Makkonen (2006, *Ref. 2*) has proved that the Gumbel Pp is the best of all.

Table 1 shows how in CumFreqA the X-values have been ranked in ascending order and the Pp values are determined. Further, the calculated Cp values have been added by fitting a probability distribution in a way that will be explained later.

Table 1. Observed and calculated cumulative probabilities

X-value Ranked	Cumulative probability (%)	
	Pp	Cp calculated
18.0	7.69	9.76
25.0	15.38	12.30
37.0	23.08	21.25
47.0	30.77	38.49
48.0	38.46	41.13
49.0	46.15	44.00
51.0	53.85	55.32
58.0	61.54	58.93
80.0	69.23	70.62
98.0	76.92	79.37
105.0	84.62	82.37
125.0	92.31	89.39

1.3. Generalizaton

The generalization is accomplished by a transformation of the data. A well known transformation is taking the logarithmic value of the data before applying the normal distribution, obtaining the log-normal distribution. When the data set is skew to the right, the normal distribution cannot be used because it is symmetrical. However, by employing the logarithmic transformation it may happen that the distribution does become normal.

In this article the transformation is realized by raising the data values to the power (exponent) E. When E < 1 the effect is similar to taking the logarithmic value. However, because the E value may have a large range its versatility is greater than only a single log transformation.

The logistic probability distribution is similar to the normal distribution. By applying the generalization, distributions both skew to the left and skew to the right can be transformed into normal or logistic distributions (*Figure 1*, *Ref. 3*).

1.4. Skewness

The generalized logistic distribution (*Ref. 3*), depending on whether the exponent E<1, E=1 or E>1 yields the results as pictured in *figure 1*.

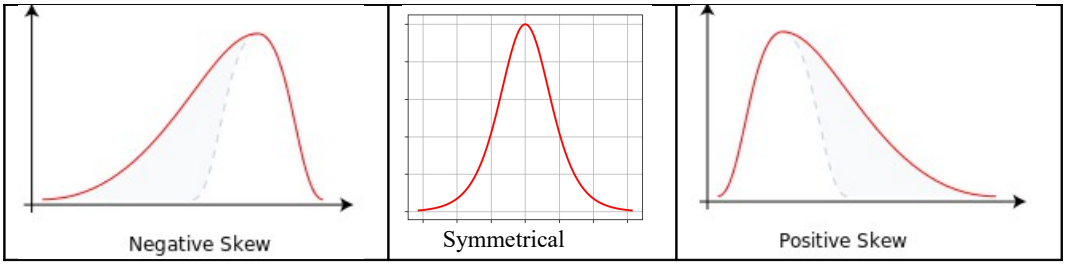

Figure. 1. *Probability density function (PDF) skewed to the left (negative skew, 1st picture), symmetrical (central picture), skewed to the right (positive skew, 3rd picture).*
With a value of E>1, the logistic distribution can transform a distribution skewed to the left into a symmetrical distribution, while with a value E<1 it can transform a distribution skewed to the right into a symmetrical distribution.

In this paper, the Gumbel and the mirrored Gumbel distributions instead of the logistic distribution. The standard logistic distribution is symmetrical, but by generalization and the value of the exponent E (see *figure 1*) it can become skewed to the left or to the right.

The standard Gumbel distribution is positively (right) skewed, but after generalization with an E value less than 1 it can become symmetrical or even left skewed.
The mirrored Gumbel distribution (see next section) is by definition left (negatively) skewed, but after its generalization with an E value greater than 1 it can approach the symmetrical normal distribution or even a right skewed one.

1.5 Mirrored distributions

The cumulative probability distribution (Pd) function can be written in general terms as:

$$Cp = f\ (X, A, B, \ldots)$$

where X is the data value, and A, B . . . are the parameters.

In of case of generalization with the use of an exponent E (*section 1.3*) the expression for Pp becomes:

$$Cp = f\ (X, A, B, E, \ldots)$$

The mirrored distribution of Cp (Cm) is simply

$$Cm = 1 - f\ (X, A, B, \ldots) \quad \text{or} \quad Cm = 1 - f\ (X, A, B, E, \ldots)$$

If Cp represents a distribution skewed to the right, then Cm will represent a distribution skewed to the left and vice versa.

Well known examples of mirrored probability distributions are the Burr and the Dagum distribution.

1.6 Linearization of the standard and mirrored Gumbel distribution

The standard Gumbel distribution can be written as:

$$Cp = \exp[-\exp\{-(A*X+B)\}]$$

where Cp is the cumulative probability distribution.

Taking the natural log (ln) of the standard Pp gives:

$$\ln(Cp) = -\exp\{-(A*X+B)\} \quad \text{or} \quad -\ln(Cp) = \exp\{-(A*X+B)\}$$

Taking the natural log once again yields:

$$\ln\{-\ln(Cp)\} = -(A*X+B) \quad \text{or} \quad -\ln\{-\ln(Cp)\} = A*X+B$$

Using the plotting position Pp, being an estimator of the cumulative probability Cp, instead of Cp and setting

$$D = B + \ln\{-\ln(Pp)\}$$

we find:

$$A*X + D = 0$$

which is the linearized form of the standard Gumbel distribution.

The parameters A and D can now be found from a linear regression so that the standard Gumbel distribution is fully defined.

For the mirrored Gumbel distribution a similar procedure can be followed. The only difference is the expression for the D value:

$$D = B + \ln\{-\ln(1-Pp)\} \quad \text{instead of} \quad D = B + \ln\{-\ln(Pp)\}$$

For the generalized forms the only difference is the expression for the linear equation:

$$A*Z + D = 0 \quad \text{with} \quad Z = X^{\wedge}E$$

The exponent E will have to be found by a numerical method maximizing the goodness of fit.

2. Overview of composition with twice the same distribution

Figure 2 gives an overview of the probability distributions used in CumFreqA (*Ref. 4*) . By clicking on a preferred distribution followed by a click on the "Confirm" box will make the program do the calculations for that particular program. However, if one selects "Best of All Distributions, the program will handle all the distributions and present the best one, but it will also give a list of rakings of all the distributions according to their goodness of fit.

The Gumbel standard, the Gumbel generalized, the mirrored Gumbel, and the mirrored Gumbel generalized probability distributions are encompassed in this overview.

Figure 2. List of probability distributions handled in CumFreq

When composed probability distributions are used, the composition consists of the introduction of a divide (a separation point or breakpoint) Xs for the X data whereby one group contains the X values smaller than Xs and the second group the X values larger than Xs.

The composition is useful when the lower X values are generated under conditions different from the higher ones. For example, the lower rainfalls occur during steady rainy periods while the higher values occur more in stormy whether.

The list of composite probability distributions included in CumFreqA is presented in *figure 3*.

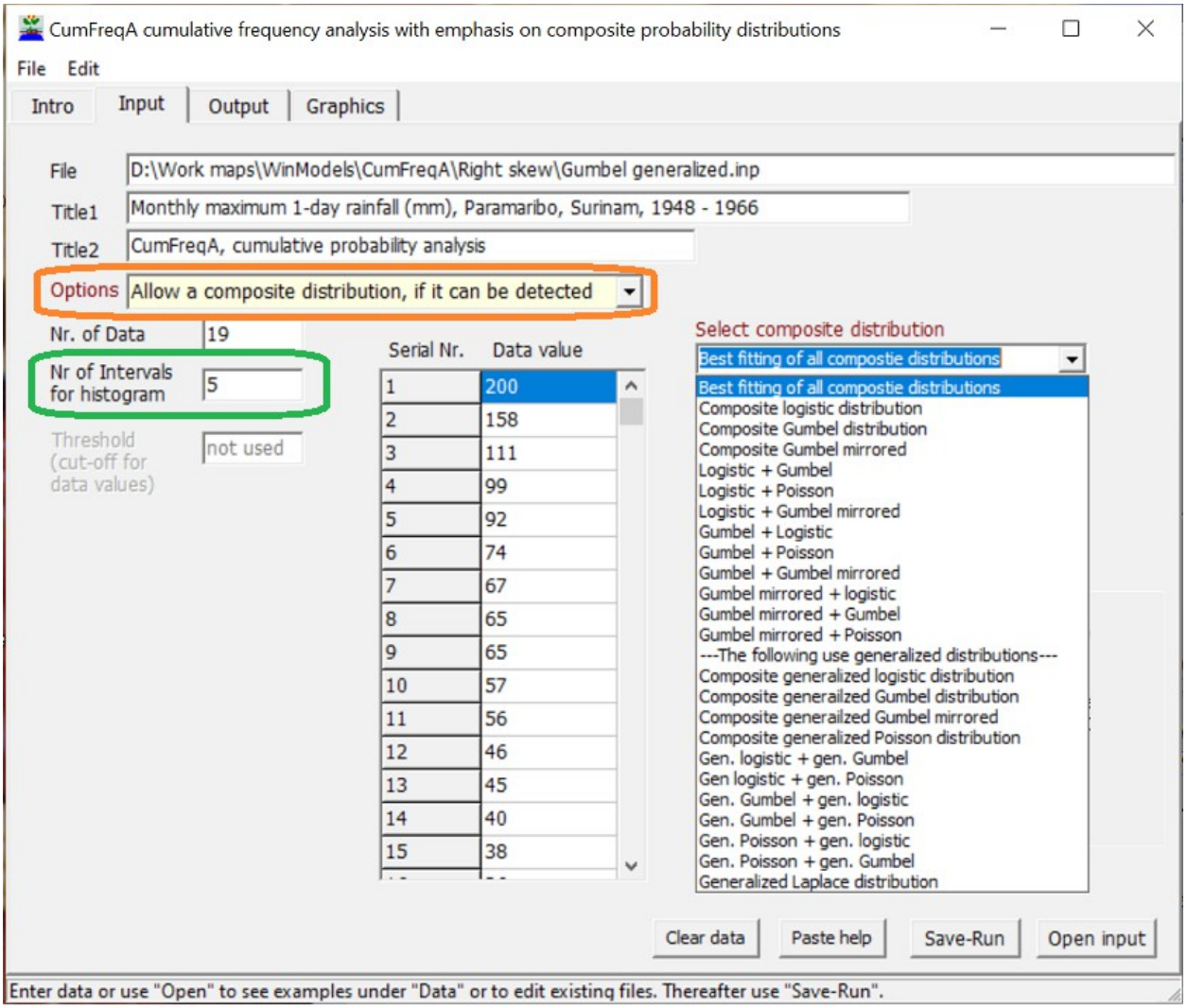

Figure 3. *List of composite distributions incorporated in CumFreqA. The composite distribution may consist of a distribution of one kind for the lower X values and a distribution of another kind for the higher X values, but also the composition can be done with one kind of distribution only whereby the parameters of the distribution for the lower values are different from those for the higher values.*

Please note that, to obtain the list, the option box (orange rectangle) must be set to "Allow a composite distribution". The green rectangle give the user the possibility to determine the number of intervals for the histogram and probability density function in the output.

The list in *Figure 3* consists of two groups. The first group works with standard probability distributions while the second group consists of generalized distributions.

In the remainder part of this paper, attention will be paid to the composite standard Gumbel distribution, the composite mirrored Gumbel distribution as well as the composite generalized Gumbel distribution and the composite generalized mirrored distribution

The composite standard Gumbel distribution can be written as:

$$Cp = \exp[-\exp\{-(A1*X+B1)\}] \quad \text{when } X < Xs$$
$$Cp = \exp[-\exp\{-(A2*X+B2)\}] \quad \text{when } X > Xs$$

The parameters A1, B1, A2 and B2 are determined in the same way as the parameters A and B for the non-composite (uniform) distribution (*Section 1.5*) , except that one works with two different data sets of the divided X values left and right of the separation point Xs.

For the mirrored and generalized distributions equations like the two previous ones can be simply formulated according to the principles explained in *Section 1.6*.

3. Examples from practice

3.1 Left skewed distribution

The data for the left skewed distribution stem from the school test scores of pupils.
As it concerns left (negatively) skewed data, the first trial will be with the mirrored Gumbel probability distribution as mentioned in *section 1.4*

As the mirrored Gumbel distribution is left skewed it will be applied first without generalization. *Figure 4* gives the cumulative probability distribution of test scores of pupils in a school in Australia fitted to a mirrored Gumbel distribution. The goodness of fit (coefficient of explanation or R-squared) is very high: 0.9976 (practically 100%). More precise versions of the mirrored distribution need not be tempted.

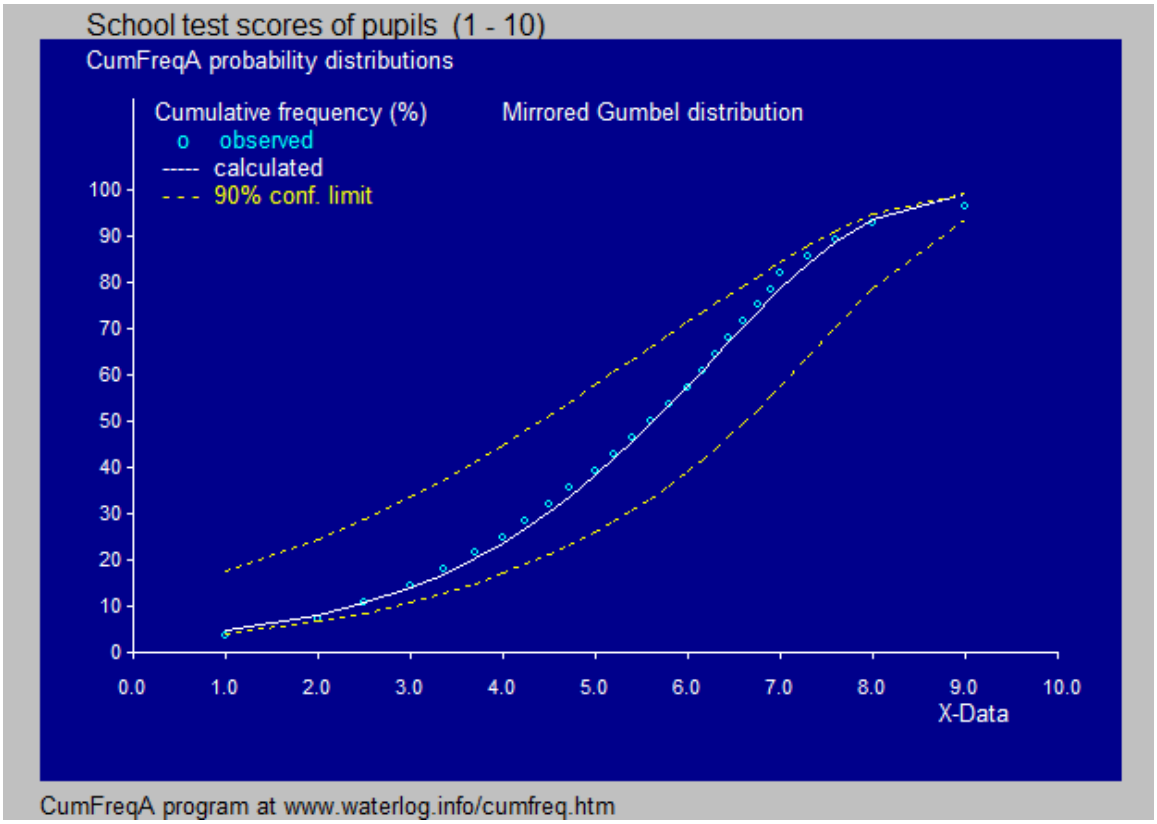

***Figure 4**. The mirrored Gumbel distribution applied to test scores.*

The proof that the distribution is skew to the left can be seen in *figure 5* in which the interval distribution and the frequency density curve are depicted.

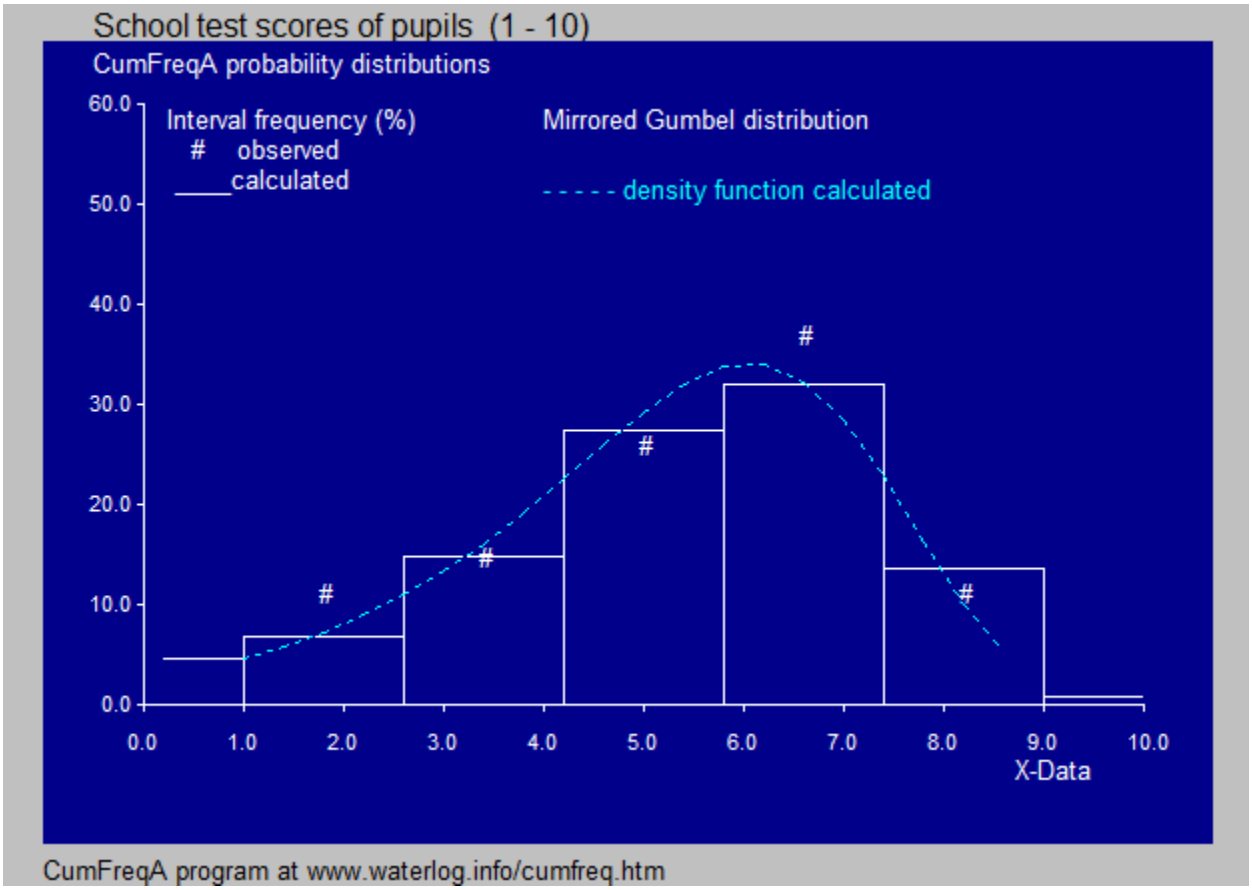

Figure 5. *Interval frequency and density curve of test scores according to the mirrored Gumbel distribution. The distribution is clearly negatively skewed (skew to the left, compare with* figure 1.

Owing the excellent fit, it is not worth to try the generalized or composite versions of the mirrored Gumbel distribution.

It could be that the standard Gumbel distribution, which originally is skew to the right, might also give a good fit when applied as the generalized version. The result of this effort is shown in *figure 6*.

The generalized Gumbel distribution can be written as:

$$Cp = \exp[-\exp\{-(A*Z+B)\}] \quad \text{with} \quad Z = X \,\hat{}\, E$$

In the case of figure 6 we have: A = 0.0170 B = - 0.972 E = 2.55
and the goodness of fit (R-squared) is 0.9958, very close to that of the mirrored version above.

Though not really required the composite generalized Gumbel distribution could be experimented. The outcome is depicted in *figure 7*.

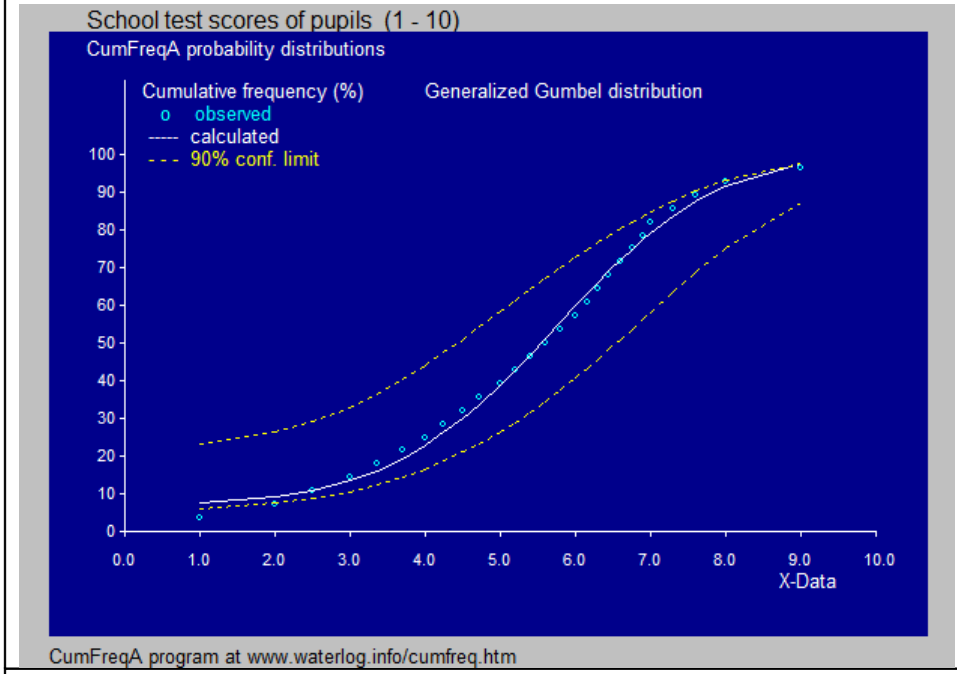

Figure 6.

*The generalized Gumbel distribution gives an equally good fit as the mirrored Gumbel distribution (*figure 5*). Owing to the fact that this generalized case gives the same result as the standard case in* figure 5*, the latter would be preferable as it has less parameters*

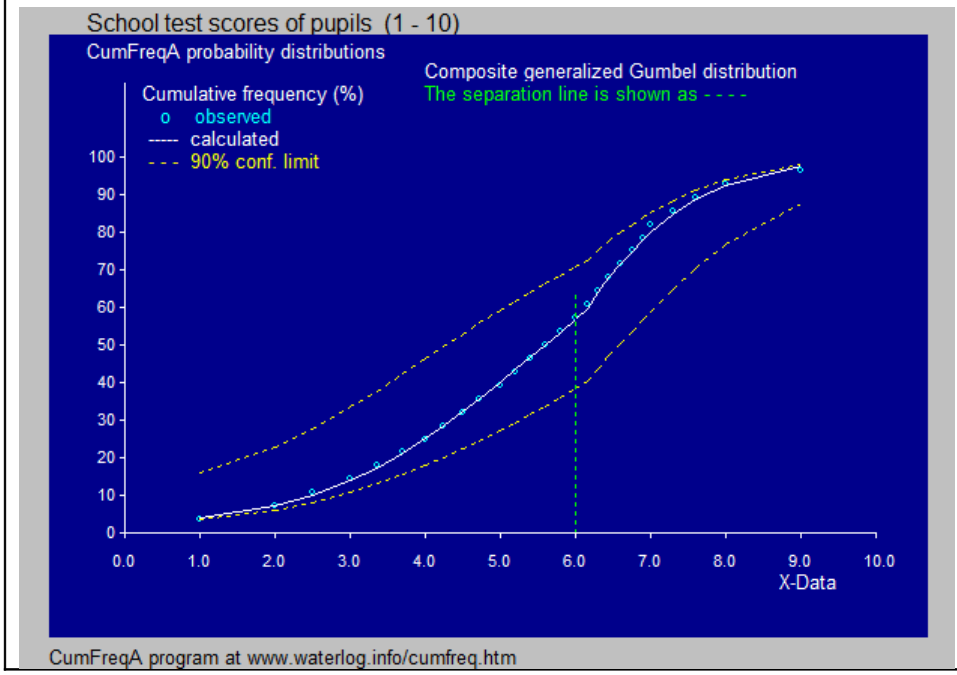

Figure 7.

The composite generalized Gumbel distribution has a separation point at Xs=0.60 (green dotted line).

The conclusion is the same as the one described in the text for figure 7.

The relevant equations are given hereunder.

With Xs=6.0 (separation point), the composite generalized Gumbel distribution (figure 7) reads:

X < Xs: Cp = exp[−exp{− (0.099*Z −1.27)}] where Z = X^1.63

X > Xs: Cp = exp[−exp{− (0.188*Z −2.19)}] where Z = X^1.62

The goodness of fit (R-squared) equals 0.9994 This is excellent but only slightly higher than in the mirrored Gumbel case (0.9976), and the difference is far from significant. However, the density function (*figure 8*) gives a somewhat better picture than that of *figure 5* as here the observed value (symbol #) is closer to the to it.

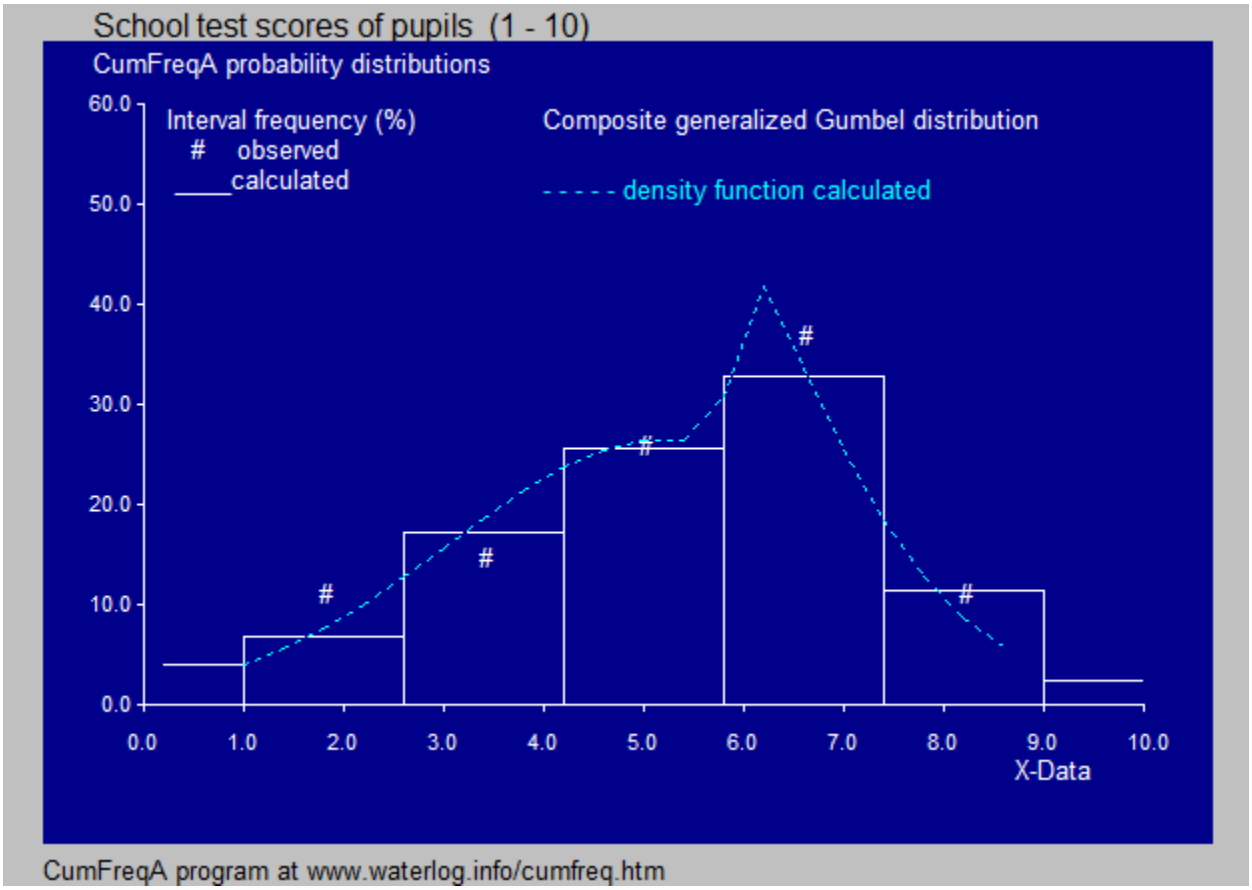

Figure 8. *Interval frequency and density function for the composite generalized Gumbel distribution showing clearly the skewness to the left.*

3.2 Symmetrical distribution

In *figure 9* the generalized Gumbel distribution is used to a dataset to a symmetrical distribution. Generalization is necessary because the standard Gumbel distribution is positively skewed (skew to the right.

The equation of the distribution in this case is:

$$Cp = \exp[-\exp\{-(0.0483*X^{1.82}-1.29)\}]$$

where the power 1.82, greater than 1, makes it possible to convert the Gumbel distribution to a symmetrical one.

In figure 10, showing the density function it can be seen that the generalized Gumbel distribution is able to produce a symmetrical distribution

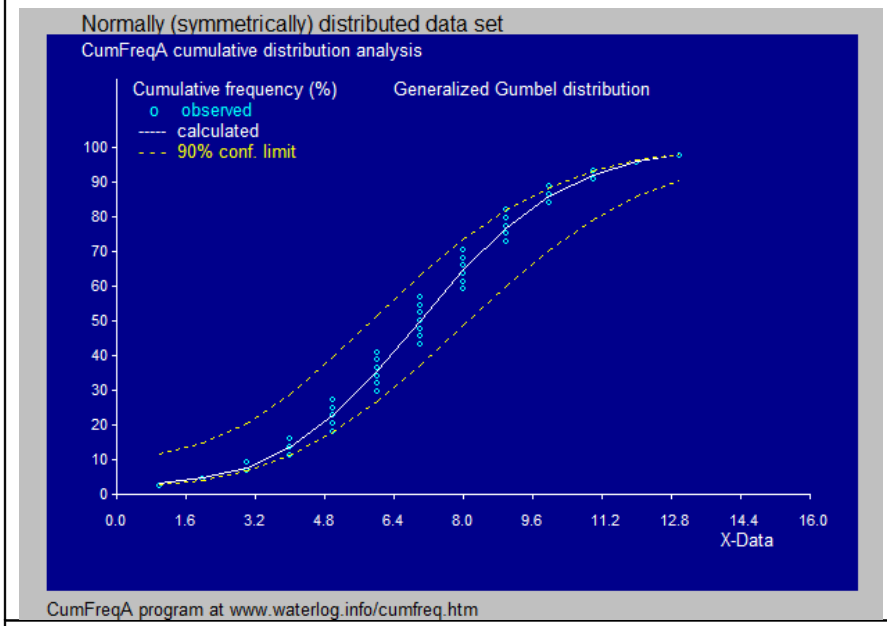

Figure 9

The generalized Gumbel cumulative probability distribution fitted to a symmetrical data set.

The goodness of fit (coefficient of explanation, R-squared) is 0.9866, close to 1 (100%);

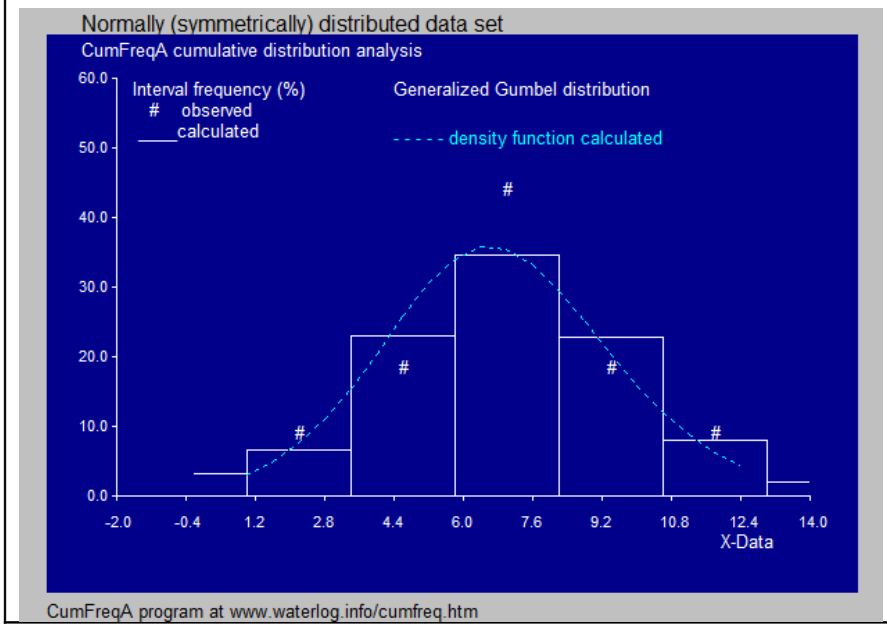

Figure 10

The eneralized Gumbel density function corresponding to the cumulative distribution in figure 9 *is clearly symmetrical.*

To approach the normal distribution the Gumbel distribution needs a generalization with an exponent E greater than 1. So, for the same purpose, the mirrored Gumbel distribution needs an exponent smaller than 1.

The equation for the generalized mirrored distribution shown in *figure 11* is:

$$Cp = \exp[-\exp\{-(-3.030*X^{0.410}-7.14)\}]$$

where the power 0.410 is smaller than 1 indeed.

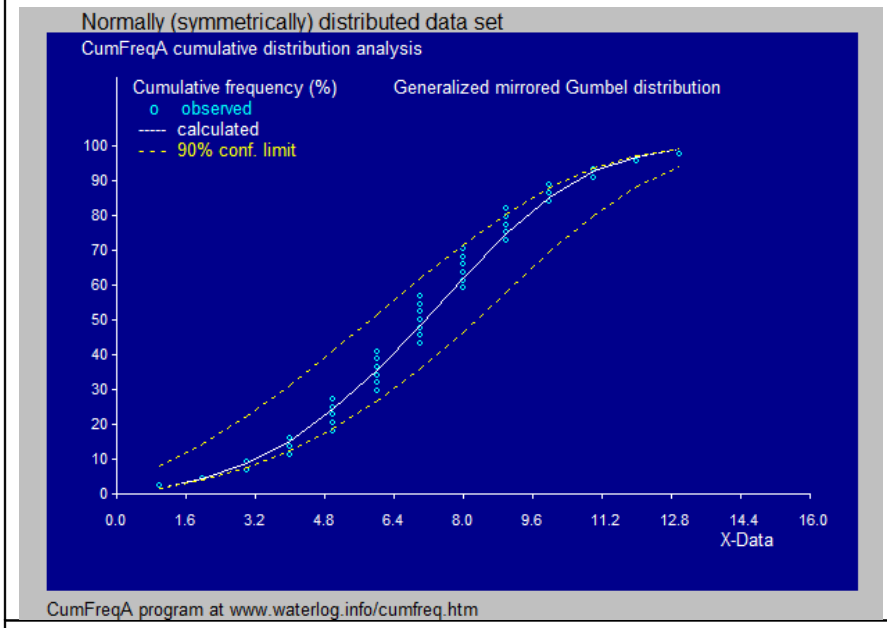

Figure 11

The generalized mirrored Gumbel cumulative probability distribution fitted to a symmetrical data set.

The goodness of fit (coefficient of explanation, R-squared) is 0.9831, *close to 1 (100%), just like the value in Figure 9.*

.

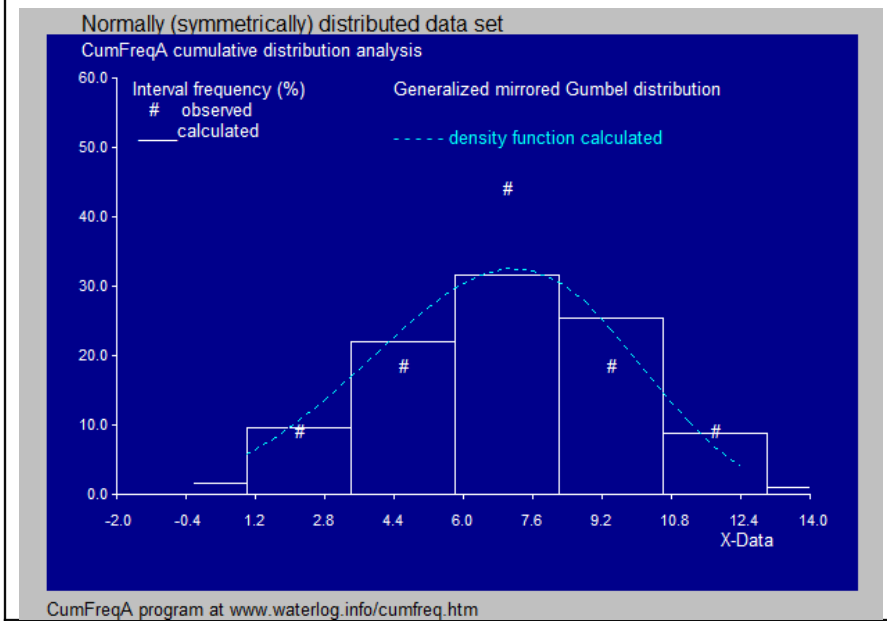

Figure 12

The generalized mirrored Gumbel density function corresponding to the cumulative distribution in figure 11 *is clearly symmetrical like* figure 10.

It can now be tried to see if the composite (discontinuous) distribution give a still better result, even though the previous results were fabulous.

Figure 13 contains the composite generalized Gumbel distribution. The corresponding equations read:

$$X < 7.0: \quad Cp = \exp[-\exp\{-(0.117*X^{1.38}-1.46)\}]$$

$$X > 7.0: \quad Cp = \exp[-\exp\{-(0.146*X^{1.41}-1.84)\}]$$

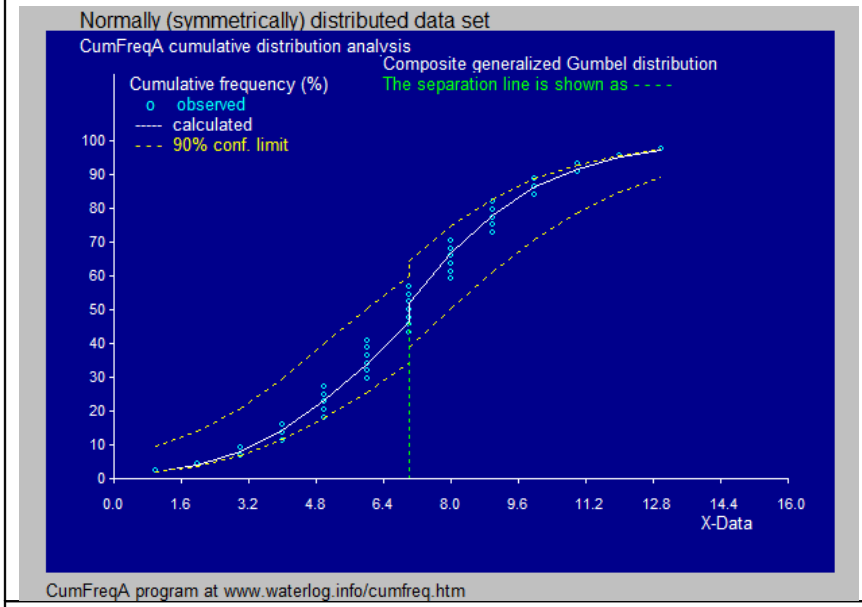

Figure 13

The composite generalized mirrored Gumbel cumulative probability distribution fitted to a symmetrical data set. The goodness of fit (coefficient of explanation, R-squared) is 0.9887*, close to 1 (100%). This value is higher than the previous ones indeed, but the improvement is very small. Moreover, there was actually not much space for improvement as the previous coefficients were already very high.*

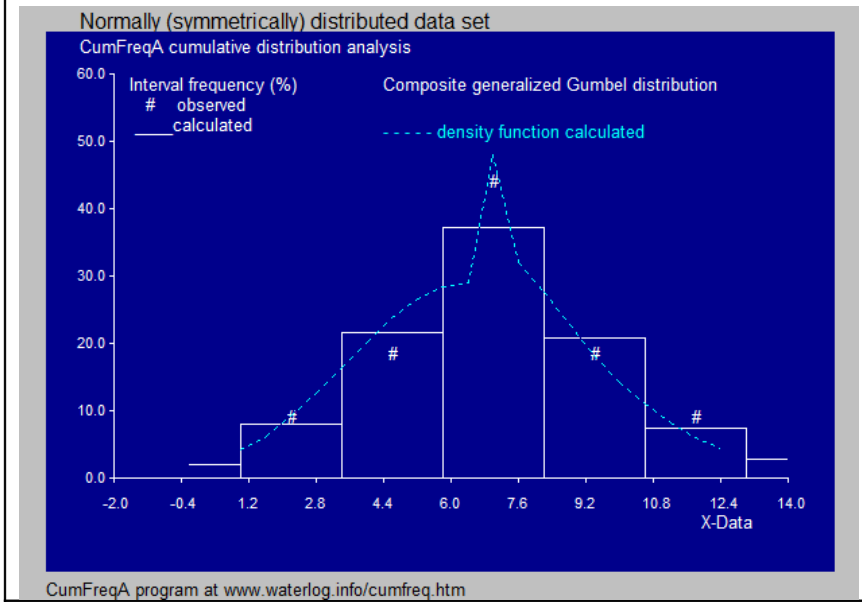

Figure 14

The composite generalized mirrored Gumbel probability density function clearly reveals the discontinuity at the separation point Xs = 7.0 And the graph approaches the maximum # symbols (observed interval frequency) nearer than in figures 10 and 12 that were made for continuous (non-composite) distributions

In the same way, it could be tried to apply the composite generalized mirrored Gumbel distribution in complementation to the composite generalized standard Gumbel distribution.

However, as the results in both cases are practically the same, the proposed option will not be further pursued.

3.3 Right skewed distribution

In this section the maximum monthly rainfalls of October in Surinam will be assessed. As the heavy rain storms occur under influence of the tropics while the more gentle rainfalls com from the Caribbean, there may be two trends in the probability curve.

First, the rainfall is analyzed without composition, but with generalization, where after it will be demonstrated that dividing the curves in two different segments with composition consisting two different versions of the same distribution type (Gumbel respectively mirrored Gumbel) will give rise to a considerable improvement, a feature that has not happened in the previous examples.

Figure 15 shows the cumulative probability according to the generalized Gumbel distribution, followed by figure 16 illustrating the density curve belonging to it.

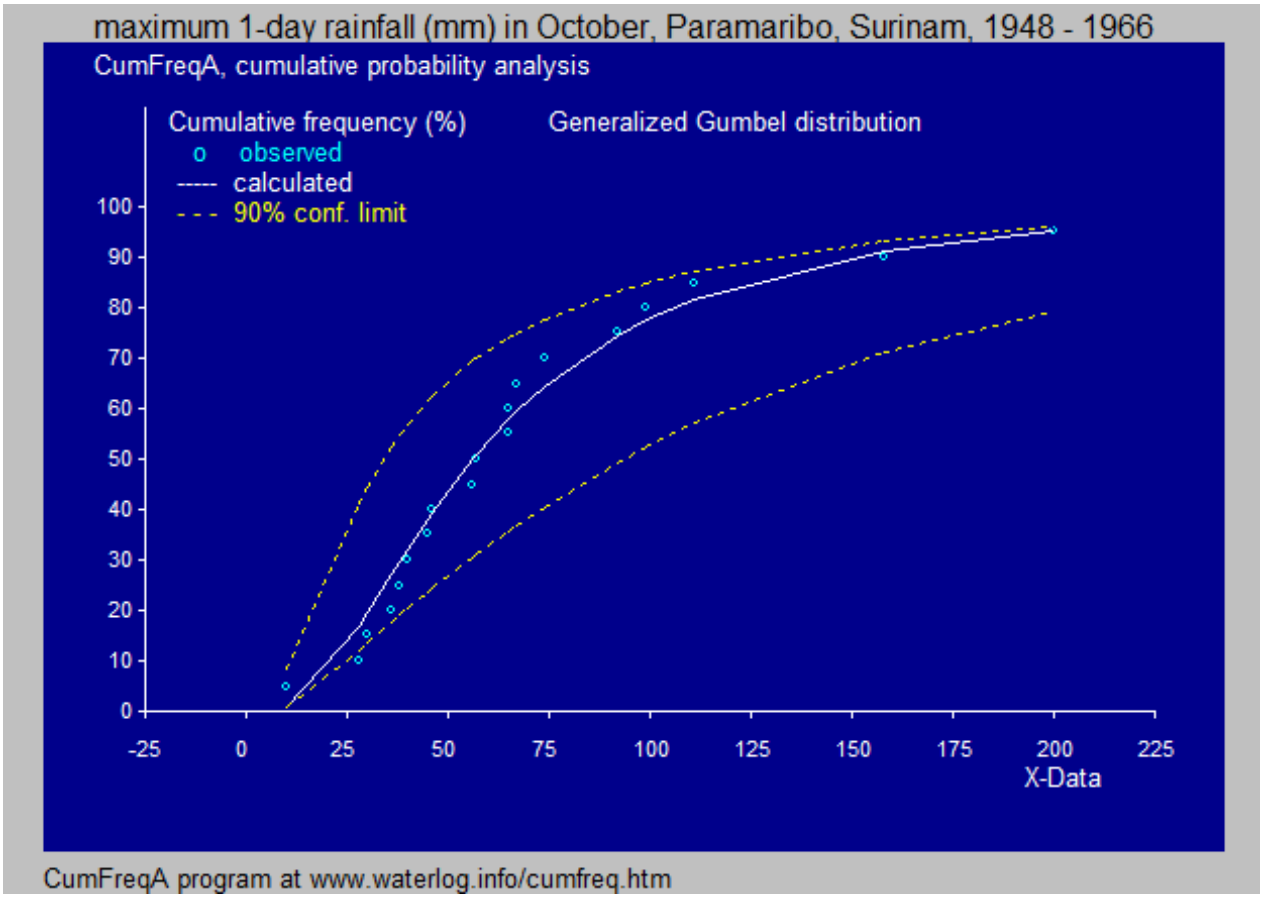

Figure 15. *The generalized Gumbel distribution applied to fit the data on the maximum 1-day October rainfall.*

The equation for the generalized mirrored distribution shown in *figure 15* is:

$$Cp = \exp[-\exp\{-(0.6850*X^{0.420}-3.35)\}]$$

and the coefficient of explanation is 0.9796

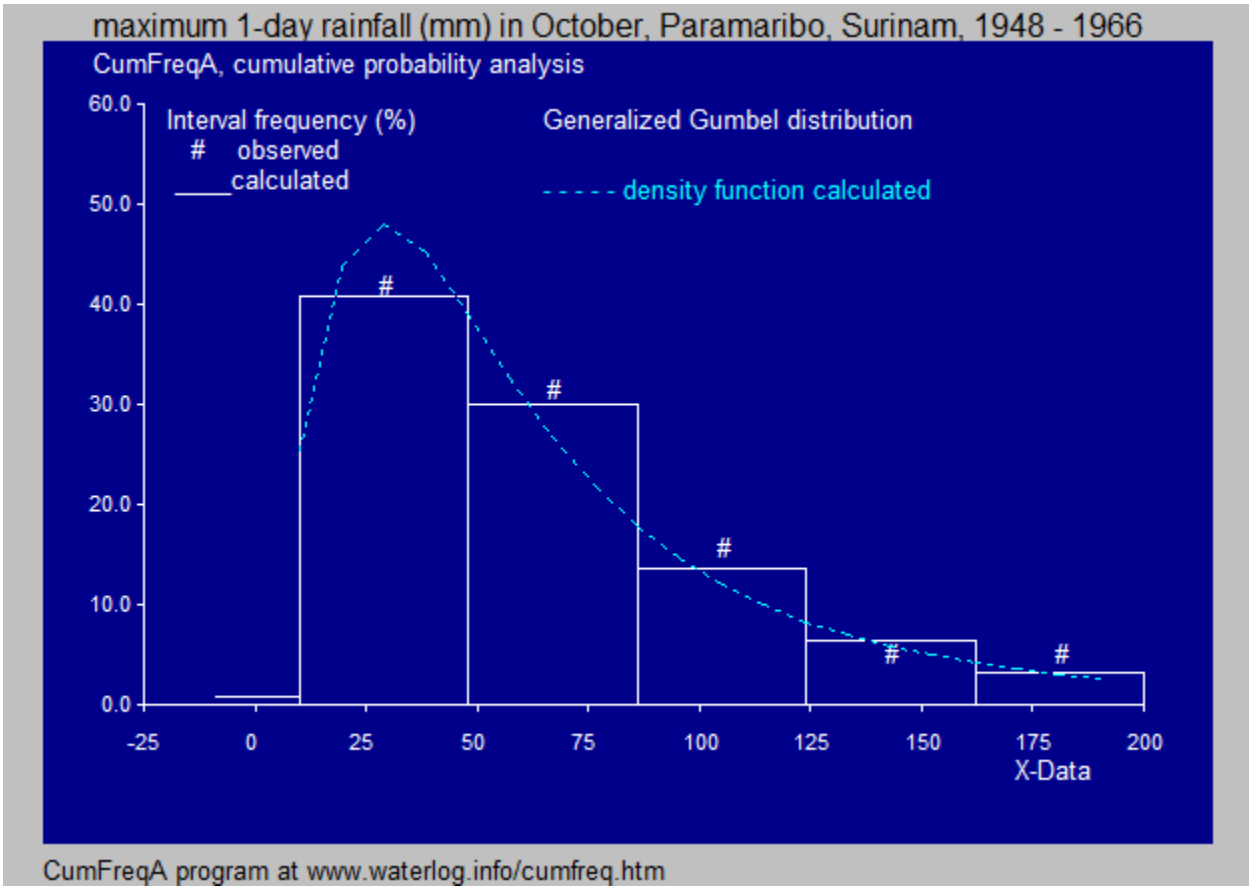

Figure 16. *The interval distribution and the density function belonging to* figure 15.

The distribution is so highly skewed to the right that the generalized Gumbel distribution, which is based on positive skewness , still needs an exponent much less than 1: E = 0.420.

Therefore it is not worth the trouble to try the generalized Gumbel mirrored distribution that is based on negative skewness, because the exponent E would be impossibly small.

The next step is to employ the composite Gumbel distribution, to see if this provides a still better result than the generalized distribution.

Figure 17 illustrates the result of this effort and *figure 18* gives the corresponding interval distribution and the density function.

$$X < 74.6: \quad Cp = \exp[-\exp\{-(0.0351*X-1.64)\}]$$
$$X > 74.6: \quad Cp = \exp[-\exp\{-(0.0160*X)\}]$$

The goodness of fit (R-squared) equals 0.9904 This is excellent but only slightly higher than in the generalized Gumbel case (0.9796), and the difference is not significant.

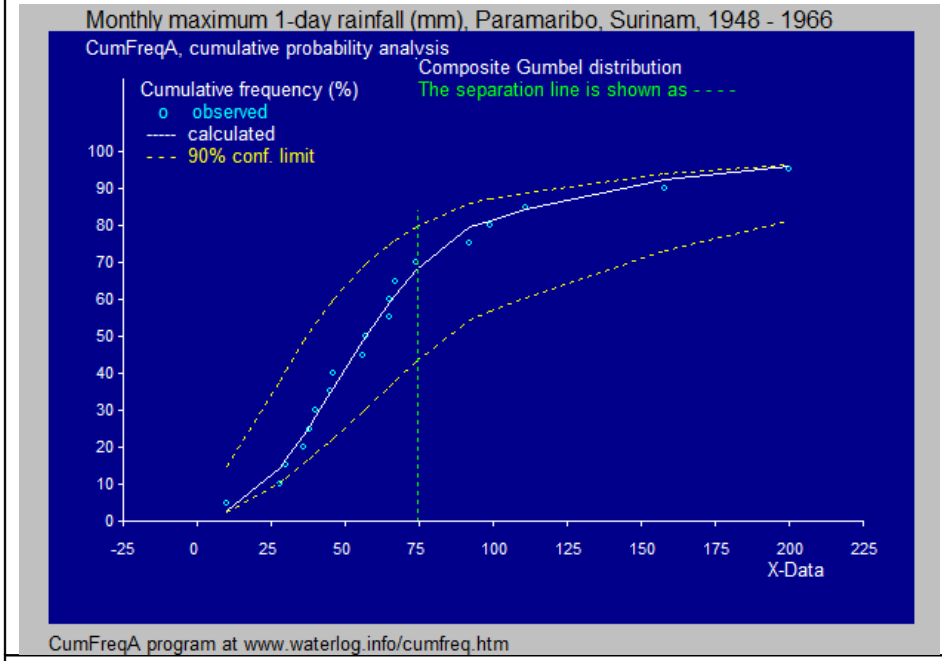

Figure 17.

Composite Gumbel distribution of monthly maximum 1-day rainfall (mm) in October, Surinam.

The separation point Xs equals 74.6 mm (green dotted line)

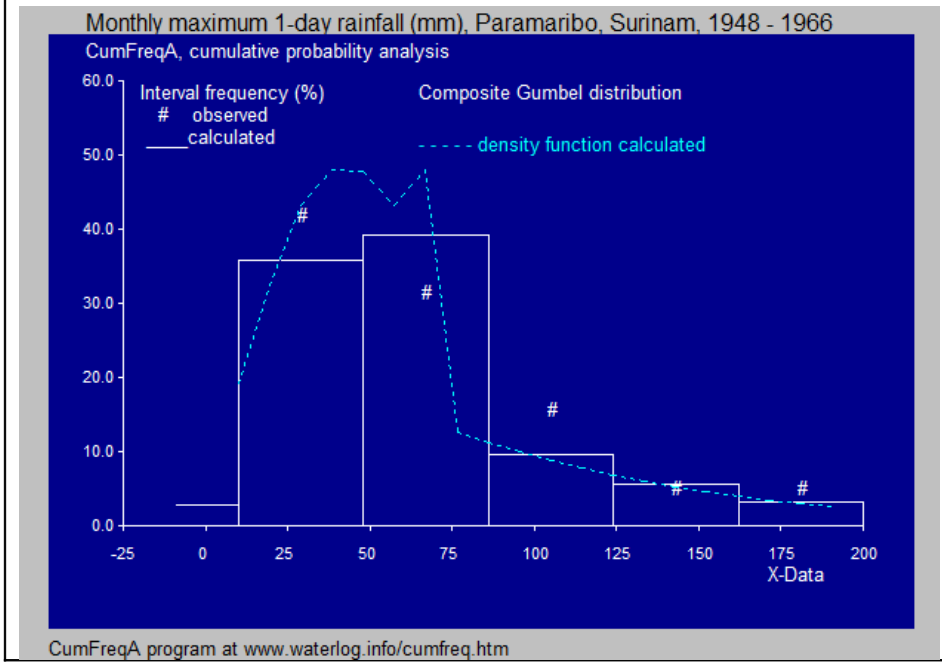

Figure 18.

Interval distribution and density function belonging to figure 17.

The breakpoint (separation point) is clearly visible.

Since the insignificant difference in the goodness of fit of the distribution shown in figure 17 with that of the one shown in figure 15, it will be tried to activate the composite generalized Gumbel distribution instead of the composite standard one.

Figure 19 illustrates the result of this effort and *figure 20* gives the corresponding interval distribution and the density function.

X < 65.0 : Cp = exp[−exp{− (0.0242*X^1.07−1.50)}]

X > 65.0 : Cp = exp[−exp{− (0.0078*X^1.13−0.009)}]

The separation point is Xs = 65.0 mm is lower than in the composite non-generalized case 74,6 (figure 17).

The goodness of fit (R-squared) equals 0. 0.9922 (very close to 1 or 100%). This is quite excellent but only slightly higher than in the generalized Gumbel case (0.9796) and in the composite standard Gumbel case (0.9904).

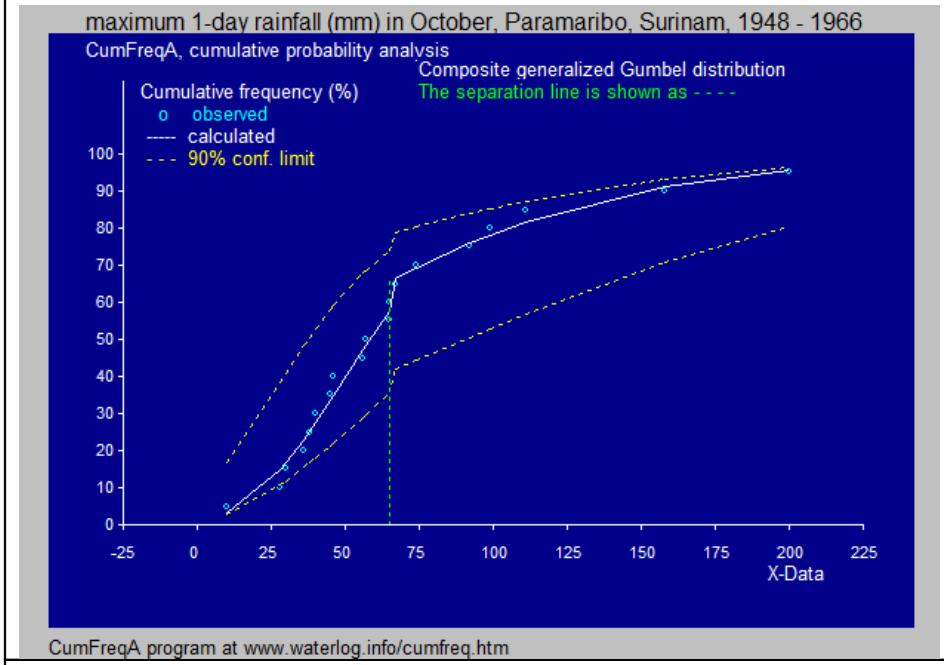

Figure 19.

Composite generalized Gumbel distribution of monthly maximum 1-day rainfall (mm) in October, Surinam.

The separation point Xs equals 65 mm (green dotted line)

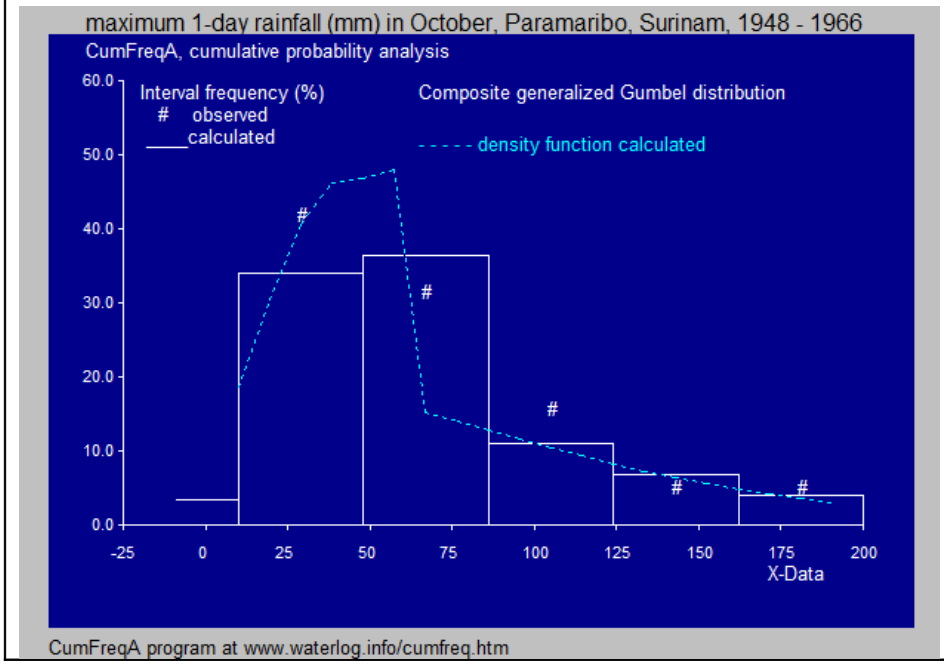

Figure 20.

Interval distribution and density function belonging to figure 17.

.

4. Conclusion

The specialty of CumFreqA to introduce inversion (mirrorization), generalization and composition to probability distribution enhances their applicability to a great extent.

This does not only hold for the Gumbel distribution discussed in this paper or for the logistic distribution (*Ref. 3*), but also for many other types of distributions.

5. References

Reference 1.
Gumbel, E.J. (1954). *Statistical theory of extreme values and some practical applications.* Applied Mathematics Series. Vol. 33 (1st ed.). U.S. Department of Commerce, National Bureau of Standards. On line: ASIN B0007DSHG4.

Reference 2.
Lasse Makkonen, 2006. *Plotting Positions in Extreme Value Analysis*. In: Journal of Applied Meteorology and Climatology Vol. 45. On line: https://journals.ametsoc.org/doi/10.1175/JAM2349.1

Reference 3.
Fitting the versatile linearized, composite, and generalized logistic probability distribution to a data set. On line: https://www.waterlog.info/pdf/logistic.pdf
or: FITTING THE VERSATILE LINEARIZED, COMPOSITE, AND GENERALIZED LOGISTIC PROBABILITY DISTRIBUTION TO A DATA SET

Reference 4.
Free CumFreqA software model for the determination of probability distributions including inversion (mirrorization), generalization and composition. On line: https://www.waterlog.info/cumfreq.htm

6. Appendix: Confidence belts

In a number of figures with the cumulative distribution depicted, their 90% confidence belts have been drawn. The confidence intervals are found from the (relative) standard deviation (Sd) of the binomial probability distribution [**Ref. A**]:

$$Sd = \mathrm{sqrt}\{Fc(1-Fc)/N\},$$

where Fc is the cumulative (non-exceedance) frequency ($0<Fc<1$), and N is the number of data.

There are only two events: Fc, the non-exceedance, or (1-Fc), the exceedance, reason why the binomial distribution is applicable.

The determination of the confidence interval of Fc makes use of Student's t-statistic (t) [**Ref A**]. Using 90% confidence limits the t-value is close to 1.7 when $N>10$.

The binomial distribution is symmetrical when Fc=0.5 (in the center of the distribution), but it becomes more skew when Fc approaches 0 or 1. Therefore Fc can be used as a weight factor in the assignation of Sd to U and L (upper and lower confidence limit respectively):

$$U = Fc + 2*1.7\ (1-Fc)\ Sd$$
$$L = Fc - 2*1.7\ Fc.Sd$$

[**Ref. A**] Use of the binomial probability distribution for confidence intervals of cumulative probability distribution functions. On line: https://www.waterlog.info/pdf/binoom.pdf