International Journal of Mathematical and Computational Methods, **4**, 1-9

SOFTWARE FOR GENERALIZED AND COMPOSITE PROBABILITY DISTRIBUTION FITTING

R.J. OOSTERBAAN
Retired from International Institute for Land Reclamation and Improvement (ILRI)
Wageningen, THE NETHERLANDS
sitemaster@waterlog.info https://www.waterlog.info

Abstract: - The use of composite probability distributions (i.e. integrated from two different distributions, including inverted/mirrored distributions) is useful in case the data sample is drawn under changing external conditions, which frequently occurs. The use of composite distributions is not widespread, with the exception of the Laplace distribution. Also, software for the fitting of such distributions to data series, with the aim to obtain an impression of the frequency of occurrence under changing external conditions, is scarce. This article uses the calculator package CumFreqA, designed for that purpose, and explains how various well known distributions can be used to obtain composite ones. Simultaneously the software finds the optimal value of the separation point Q of the different distributions left and right of it. Another technique of CumFreqA is to raise the data to a power P, whose value can be optimized numerically using iterative procedures, to reach the condition of minimum sum of squares of deviations of the theoretical from the observed values. The transformation of data to obtain a better fit is not often done with the exception of the log-normal distribution which uses a logarithmic transformation of the data instead of an exponential, which offers more flexibility. Thus the distribution is generalized and made composite to enhance the goodness of fit. Further, the parameters of the distributions are found from transformations of the cumulative distribution function leading to linear equations where after a linear regression is applied, which simplifies the algorithm. The confidence belts of the cumulative distribution functions in CumFreqA are constructed with the help of the binomial distribution. This leads to the possibility to construct confidence intervals of the return period as well. Various examples of distributions and confidence belts are given. CumFreqA offers the possibility to create histograms with intervals by choice and constructs the corresponding probability density functions, of which examples are given.

Key words: - Probability distribution fitting, composite, generalization, transformation, linear regression, confidence belt

1 Introduction, Methods Used

The CumfreqA calculator package has been designed to fit composite probability distributions to data series obtained under changing external conditions. For example, the rainfall in Northern Peru follows a different pattern when the Pacific Ocean current El Niño has descended down from Ecuador compared to the situation when the current has retreated form the Peruvian coast back to Ecuador or even Colombia, and a distribution integrated from two components then provides more realistic results.

The software uses continuous cumulative probability functions (CDF) that can be transformed and linearized, such as the Burr, Cauchy, Dagum, Exponential (Poisson type), Fisher-Tippet type III (F-T III), Frechet (F-T II), Generalized Extreme Value (GEV), Gompertz, Gumbel (F-T I), Kumaraswamy, Laplace, Logistic, Student's t (with 1 and 2 degrees of freedom), Pareto-Lomax and the Weibull distribution [Ref. 1]. Although the normal distribution has no explicit expression for the CDF, it is still included in the model using the numerical approximation of Hastings [Ref. 2]

To expand the scope, the distributions mentioned, in so far they are skewed (i.e. not symmetrical), are also used in their inverted (mirrored) form, so that a distribution that is skewed to the right becomes a distribution skewed to the left, and vice versa.

The aim of composition is to find two different probability distributions, one left and one right of a separation point (Q), so that the different patterns of the data can be caught with a higher degree of goodness of fit. The two components may consist of children of the same mother distribution, but also of different mothers. Thus a composite distribution is obtained. In statistics, composite distributions are scarce, although the composite Laplace distribution is well known [Ref. 3].

The composite Weibull-Gamma distribution has also been used, but it needs an R package [Ref. 4].

The composite distribution is also known as a two-component spliced distribution [Ref.5].

The CumFreqA model uses two parameters for probability distribution fitting that are to be found by numerical optimization with the condition of Least Sum of Squares of Deviations (LSSD) of the theoretical from the observed values. These are the power (exponent) P to which the data are raised (0.2<P<3) and the separation point (Q) of the data. In statistics, the exponential transformation of the data is not common. More often, logarithmic transformations have been done as in the log-normal distribution [Ref. 6]. The exponential transformation, however, is more versatile.

The parameters of the distributions are found from a linear regression analysis of the transformed and thus linearized CDF functions.

2 Generalization and Linearization

The generalization of a symmetrical distribution like the normal and logistic distribution by raising the data values to an exponent P will result in a distribution skewed to the left when P<1 and to the right when P>1 (Figure 1).

The continuous probability distributions mentioned in the introduction are generalized and linearized as shown in Table 1. The symbols used here are explained in the frame hereunder:

Symbols used in Table 1

Fc = cumulative probability *), X = stochastic variable, A = distribution parameter,
B = distribution parameter, C= distribution parameter to be optimized numerically,
E = exponent, P = power for generalization to be optimized numerically, Ft = transformed Fc,
Z = X^P, the generalized X, Xt = transformed X, Zt = transformed Z ,
^ = raised to the power P or exponent E, * = multiplication, / = division,
Sr(Y) = square root of Y, Y = a variable, pi = 3.141….,
Ln(Y) = natural logarithm of Y (with base e), e = base of LN = 2.71 . . . . , Exp(y) = e^y.

The Fc values are estimated as Fc = R/(N+1), i.e. the plotting position, where:
R = rank number in an ascending order of Xi (i=1 . . N) and N = number of data [Ref. 7]

Table 1. Overview of (inverted/mirrored) continuous probability distributions, their generalization, transformation and linearization as used in CumFreqA

Distribution (alphabetically)	Cumulative Distribution Function (CDF)	Transformation and linearization	Comments
Burr, original (Dagum mirrored ^)	Fc = 1 - [B/(X^A+B)]^E	Xt = Ln[B/(X^A+B)] Ft = Ln(1-Fc) Ft = E*Xt	B>0, X>B, B is to be optimized iteratively Use ratio method *) to find E
Cauchy generalized #)	Fc = (1/pi)*arctan(A*Z+B) + 0.5	Xt = Z =X^P Ft = tan{pi*(Fc-0.5)} Ft = A*Xt + B	Use linear regression to find A and B
Dagum, original (Burr mirrored) ^)	Fc = [B/(X^A+B)]^E	Xt = Ln[B/(X^A+B)] Ft = Ln(Fc) Ft = E*Xt	B>0, X>B, B is to be optimized iteratively Use ratio method *) to find E
Exponential generalized	Fc = 1 – Exp{-(A*Z+B)}	Xt = Ln(Z) = P*Ln(X) Ft = - Ln (1-Fc) Ft = A*Xt + B	Use linear regression to find A and B
Generalized Exponential mirrored ^)	Fc = Exp{-(A*Z+B)}	Xt = Ln(Z) = P*Ln(X) Ft = -Ln (Fc) Ft = A*Xt + B	Use linear regression to find A and B
Fisher-Tippet III (original)	Fc = Exp[-{(C-X)/Exp(-B/A)}^A]	Xt = Ln(C-X) Ft = Ln{-Ln(Fc)} Ft = A*Xt + B	X<C, C is to be optimized iteratively. Use linear regression to find A and B
Fisher-Tippet III (mirrored)^ ^)	Fc= 1 - Exp[-{(C-X)/Exp(-B/A)}^A]		

Table 1 continued on next page

Table 1. Continued (for explanation of symbols see the frame on the previous page)

Distribution (alphabetically)	Cumulative Distribution Function (CDF)	Transformation and linearization	Comments
Frechet (F-T II) (original)	Fc = Exp[-{(X-C)/Exp(-B/A)}^A]	Xt = Ln(X-C) Ft = Ln{-Ln(Fc)} Ft = A*Xt + B	X<C, C is to be optimized iteratively. Use linear regression to find A and B
GEV #) (skew to right)	Fc = Exp[-{1+C(X-A)/B}^(-1/C)]	No transformations	A, B and C are to be optimized numerically
Gompertz Generalized #)	Fc = 1 – exp[A*{exp(B*Z) - 1}]	Xt = exp(B*Z) – 1 Ft = Ln (1-F) Ft = A*Xt	B and P are to be optimized numerically . Use ratio method *) to find A
Gumbel (F-T I) Generalized	Fc = Exp[-Exp{-(AZ+B)}]	Xt = Ln(Z) = P*Ln(X) Ft = -Ln{-Ln(Fc)} Ft = A*Xt + B	Use linear regression to find A and B
Gumbel generalized, inverted (mirrored) ^)	Fc = 1 – Exp{-Exp{-A*Z+B)}	Xt = Ln(Z) = P*Ln(X) Ft = -Ln{-Ln(1-Fc)} Ft = A*Xt + B	Use linear regression to find A and B numerically
Kumaraswamy original #)	Fc = 1 - {1 - (X/C)^B}^A	Xt = Ln{(X/C)^B} = B*Ln(X/C) Ft = Ln(1-Fc) Ft = A*Xt	C>Xmax, to be optimized numerically . Use ratio method *) to find A
Laplace, composite, generalized	X<Q: Fc = 0.5*Exp{A1*(X^P1-B)} X>Q: Fc = 1 - 0.5*exp{A2*(X^P2-B)	X<Q: Ft = ln(2Fc) C = -A1*Q Ft = A1*X^P1 + C X>Q : Ft = ln(0.5) - ln(1-Fc) Ft = A2*X^P2	C and A1 are found from a linear regression A2 comes from the ratio method *)
Logistic generalized (any skewness)	Fc = 1/(1+Exp(A*X^E+B)	Xt = Ln(X^E) = E*Ln(X) Ft = Ln(-1+1/Fc) Ft = A*Xt + B	Use linear regression to find A and B
Normal generalized (any skewness)	No analytical equation available. Hastings numerical approximation is used	Y = 1/(1+0.232X^P) N = {1/Sr(2pi)}Exp(-X^2/2) Fc = 1 - N(1 0.319 Z – 0.357 Y^2 + 1.781 Y^3 - 1.821 Y^4 + 1.330Y^5)	
Student (1 d.f.) (symmetrical)	Fc = 0.5 + arctan{(X-AvX)/StD}/pi AvX= Average of X StD = Standard deviation of X	No transformation	
,Student (2 d.f.) (symmetrical)	Fc = 0.5{1+(RedX)}/Sr(2+RedX^2) RedX = (X-AvX)/StD		
Pareto-Lomax #)	Fc = 1 - {B/(X+B)}^E	Xt = Ln{B/(X+B)} Ft = Ln(1-Fc) Ft = E*Xt	B>0, X>B, B is to be optimized Use ratio method *) to find E
Weibull generalized	Fc = 1 - Exp{-(Z/C)^A} with C = Exp(-B/A)	Xt = Ln{Ln(Z)} Ft = Ln{-Ln(1-Fc)} Bt = B/A Ft = A*Xt + Bt	A and Bt are found from a linear regression
Weibull generalized mirrored ^)	Fc = Exp{-(Z/C)^A} with C = Exp(-B/A		

*) The ratio method is a linear regression while forcing the line to go through the origin..

#) These distributions can also be mirrored.

^) For mirrored (inverted) distributions see figure 1

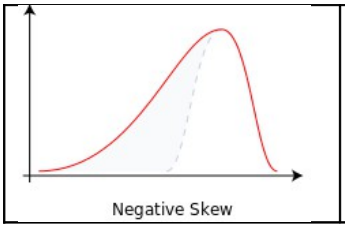
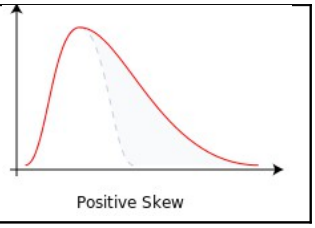


Fig.1. A probability density function (PDF) skewed to the left (1st picture) becomes skewed to the right (2nd picture) after being inverted (mirrored) and vice versa.



3 Example, Generalized Distribution, No Composition Needed

Figure 2 depicts the fitting of a generalized non-composite probability distribution. The figure is given to illustrate the facilities of the CumFreq calculator and the advantages of generalization. In figure 2 the goodness of fit is 99.1%. In this case it would not be worth the trouble to go for a composite distribution even though it might yield a still better fit, but the difference cannot be significant.

Figure3 shows the histogram and probability density function pertaining to Figure 2.

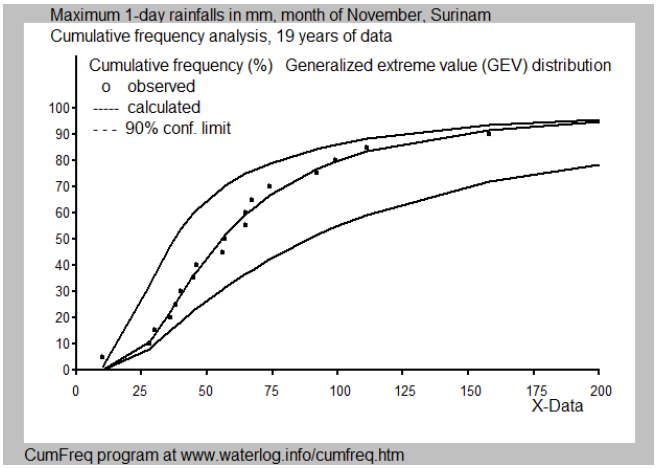


Fig.2 The best fitting non-composite probability distribution (GEV) to maximum 1-day rainfalls, month of November, Suriname [Ref. 8].

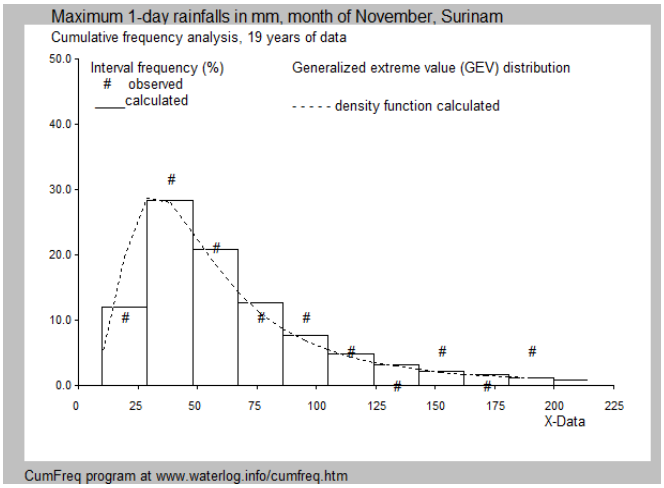


Fig.3 Histogram and probability density function for the distribution depicted in Fig. 2.

When requesting the best fitting of all distributions, CumFreq produces a list with rankings of all distributions by goodness of fit (Figure 4).

In Figure 4 the distribution shown in Figure 2, made for maximum 1-day rainfalls in the months of November, Suriname [6], ranks first.

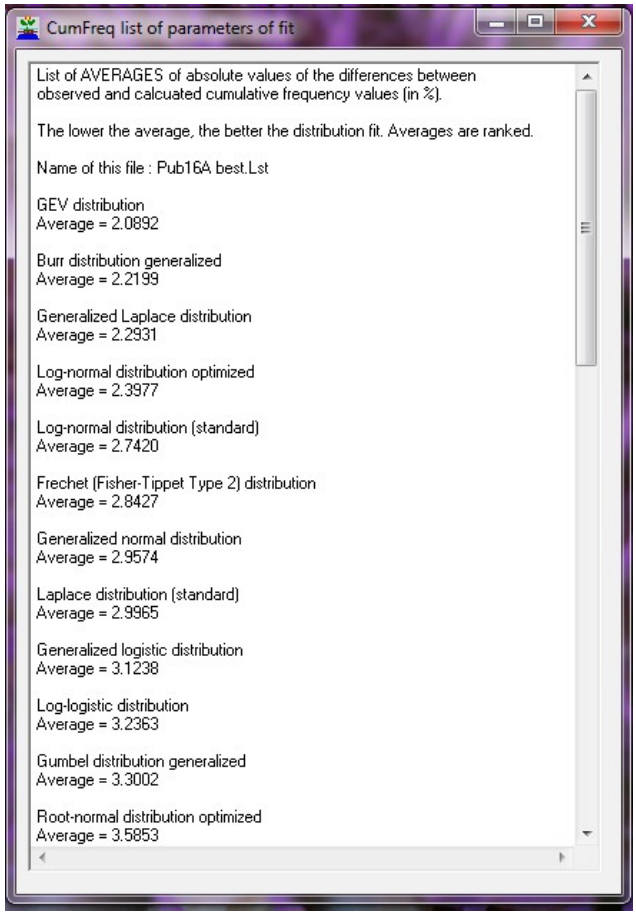


Fig.4 List of probability distributions ranked by goodness of fit. The list includes standard and generalized distributions. Only the top-part of the list is shown and the generalized distributions dominate as they give better fits.

The return periods of the runoffs shown in Figure 2, together with their confidence intervals, are as shown in Figure 5. At higher rainfalls the confidence intervals become wide so that those return periods are not robust.

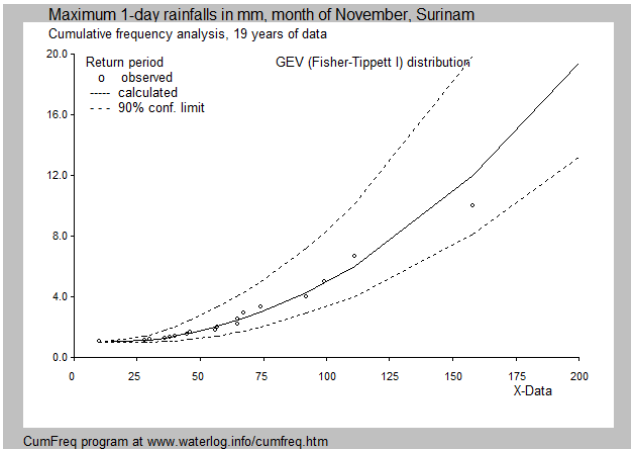


Fig.5 Return periods of rainfalls with 90% confidence limits

4 Examples, Composition Needed

Using CumFreqA [Ref. 9], examples will be given with the November rainfalls measured in Paramaribo, Suriname, during 19 years (1948-1966) [Ref. 8]. The standard Gumbel (P=1), generalized Gumbel (P>1 or P<1), composite Gumbel P1=P2=1), and composite generalized Gumbel distributions are used. These distributions are not the best of all for the data (see Figure 4), but they give a clear illustration of the principles (Figure 6).

Figures 6.1, 6.2, 6.3 and 6.4 illustrate that the generalized distributions perform better than the standard ones (with Power P=1) and that the composite distributions have a better fit than the singular ones.

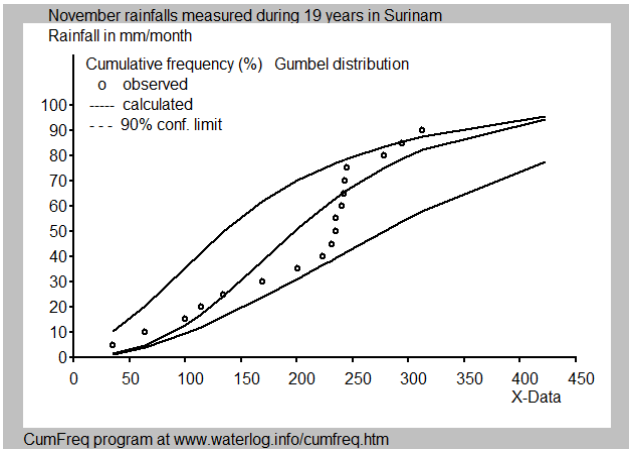


Fig.6.1 Standard Gumbel distribution Goodness of fit 89%

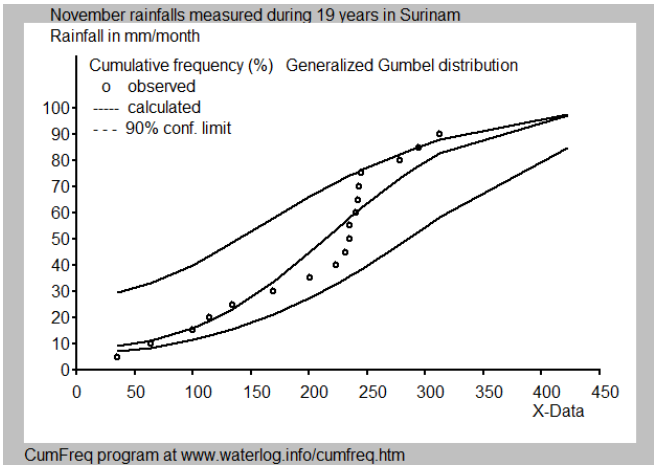


Fig.6.2 Generalized Gumbel distribution. Goodness of fit 93%

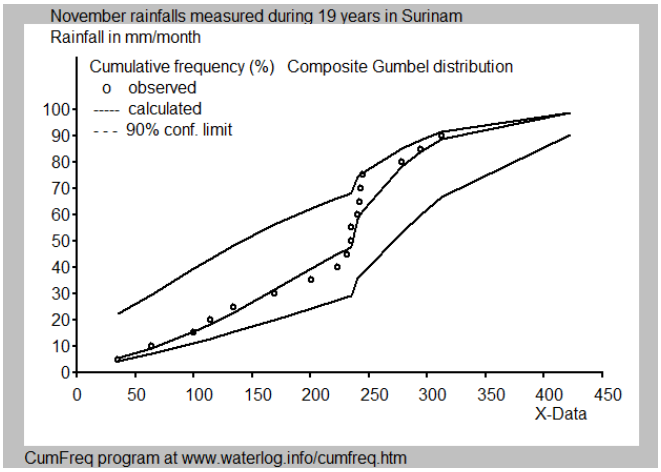


Fig.6.3 Composite standard Gumbel distribution. Goodness of fit 96%

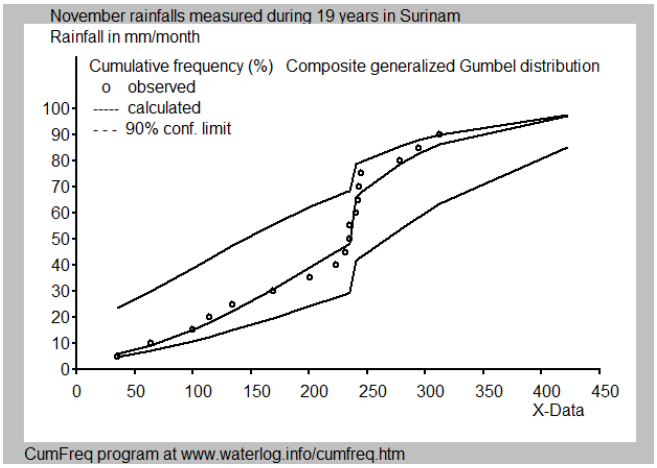


Fig.6.4 Composite generalized Gumbel distribution. Goodness of fit 98%

Fig.6 (6.1, 6.2, 6.3 and 6.4) Illustration of goodness of fit of various forms of the Gumbel distribution to the rainfall data in Surinam.

Table 2 gives the equations and their parameter values for the four cases shown in Figure 6 as they appear in the CumFreqA output. When a generalized distribution is used the values of the exponent P of the X-data is given. When a composite distribution is used, the value of the separation point Q is also given.

Table 2. CumFreqA output data showing distribution equations and parameter values for the four cases shown in Figure 6**.**

Output Standard Gumbel distribution	**Output Generalized Gumbel distribution**
RESULTS OF CumFreqA CALCULATOR Cumulative frequency analysis November rainfalls measured during 19 years in Suriname Rainfall in mm/month Name of input file used: C:\SegRegA\Gumbel standard.inp Number of data used: 19 Probability distribution preferred by user. The cumulative frequency function is double exponential (Gumbel): Fc = exp[-exp{-(A*X+B)}] A = 0.0111 B = -1.83 Mode = 1.6461E+002 The standard error of X is optimized as 114.84	RESULTS OF CumFreqA CALCULATOR Cumulative frequency analysis November rainfalls measured during 19 years in Suriname Rainfall in mm/month Name of input file used: C:\SegRegA\Gumbel generalized.inp Number of data used: 19 Probability distribution preferred by user. Cumulative frequency function of the generalized Gumbel type : Fc = exp[-exp{-(A*X^P+B)}] The exponent P = 1.84E+000 A = 6.6615E-005 B = -0.921
Output Composite Standard Gumbel distribution	**Output Composite Generalized Gumbel distribution**
RESULTS OF CumFreqA CALCULATOR Cumulative frequency analysis November rainfalls measured during 19 years in Suriname Rainfall in mm/month Name of input file used: C:\SegRegA\Gumbel composite.inp Number of data used: 19 Probability distribution preferred by user. The cumulative frequency function is composite Gumbel : The separation point is : Q = 236.240 X < Q : Freq = exp[-exp{-(As*X+Bs)}] As = 0.00687 Bs = -1.31 X > Q : Fc = exp[-exp{-(Ag*X+Bg)}] Ag = 0.0207 Bg = -4.36	RESULTS OF CumFreqA CALCULATOR Cumulative frequency analysis November rainfalls measured during 19 years in Suriname Rainfall in mm/month Name of input file used: C:\CumFreqA\Gumbel generalizedcomposite.inp Number of data used: 19 Probability distribution preferred by user. The cumulative frequency function is composite Gumbel generalized: The separation point is Q = : 235.000 X < Q : Freq = exp[-exp{-(As*X^Ps+Bs)}] As = 0.003 Bs = -1.204 The value of exponent Ps is: 1.15E+000 X > Q : Fc = exp[-exp{-(Ag*X^Pg+Bg)}] Ag = 0.005 Bg = -2.050 The value of exponent Pg is: 1.17E+000

Figure 7 shows the ranking list of standard, generalized and composite distributions for the case under discussion (Suriname November rainfall data) which is part of the CumFreqA output when the option "Best of All" is selected with the input. It is seen that the composite Logistic+Poisson (or exponential) distribution ranks first and that the composite generalized Gumbel distribution (Figure 6.4) ranks 5, while the composite standard Gumbel distribution (Figure 6.3) ranks 11.

It is also seen that the composite distributions are the highest ranking illustrating the need of a composite distribution instead of a non-composite one as in Figures 2 and 3.

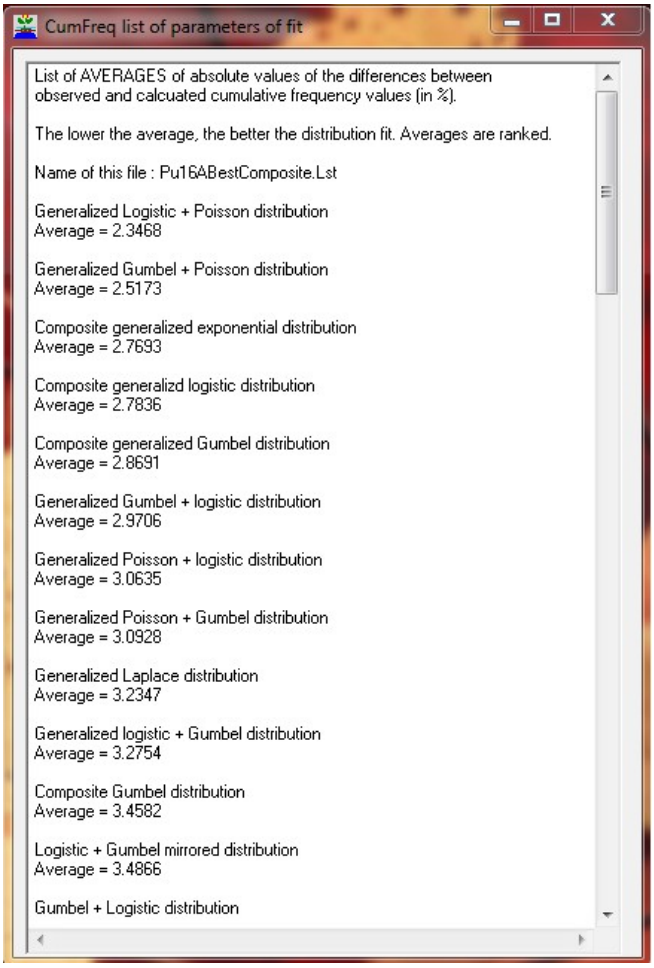


Fig.7. Ranking list of probability distributions for the Suriname data, as a part of SegRegA output. The list includes generalized, mirrored, and composite distributions. Only the top-part of the list is shown exhibiting mainly composite distributions owing to the pattern change of the plotting positions around X=235

5 Confidence Belts

In Figure 6 the 90% confidence belts of the CDF's have been drawn. The confidence intervals are found from the (relative) standard deviation (Sd) of the binomial probability distribution [Ref. 10]:

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

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 11]. 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):

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

6 Histograms and Probability Density Functions (PDF)

CumFreqA has the possibility to make histograms and PDF's [Ref. 12] with a number of intervals that can be selected by the user. Figure 8 illustrates this for the generalized composite Gumbel distribution shown in Figure 6.4 using 5 and 10 intervals. The PDF's are obtained by differentiation of the CDF. They correspond well with the histograms with an exception at the separation point. The discontinuity at the separation point Q = 235 (Table 2 right under), where the left hand distribution changes into the right hand distribution, is clearly visible. The distance between the observed interval frequency and the calculated (theoretical) frequency appears to be relatively small, except at X=Q.

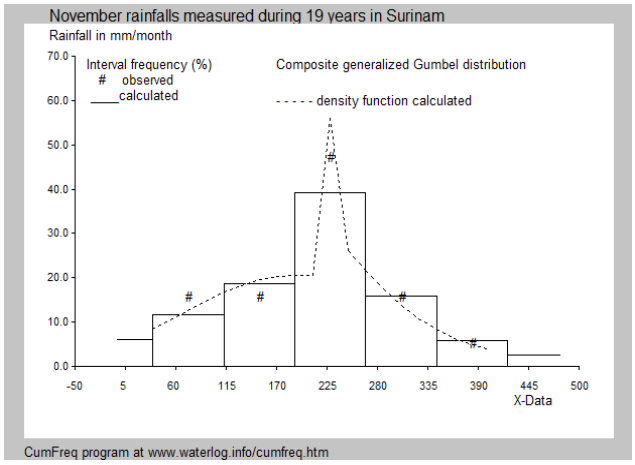


Fig.8.1 Histogram with 5 intervals

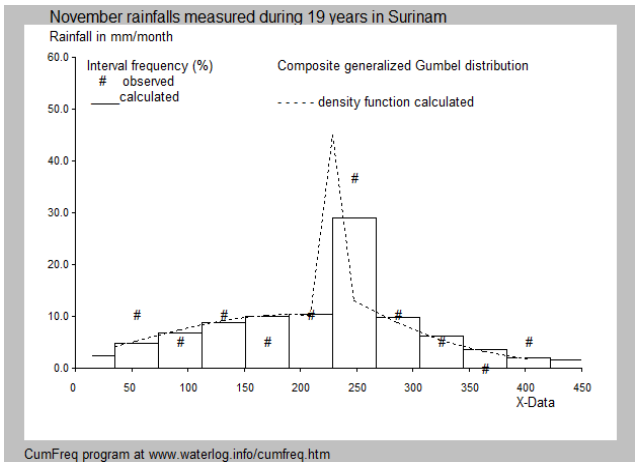


Fig.8.2 Histogram with 10 intervals

Fig.8. CumFreqA histograms and corresponding probability density functions for the composite generalized Gumbel distribution of Fig. 6.4.

7 Return Periods

For data obtained in a time sequence, like the rainfall data under study, the return period is a much used characteristic. It is defined as T = 1/(1-Fc) and is expressed in time units. For the yearly November rainfalls, the unit is year. CumFreqA produces tables and graphs of return periods together with their confidence belts (Figure 9). In this figure it can be seen that a November rainfall of 300 mm or more is estimated to return on average every 7 years, but there is a 90% chance that the actual return period is somewhere between 4 and 10 years.

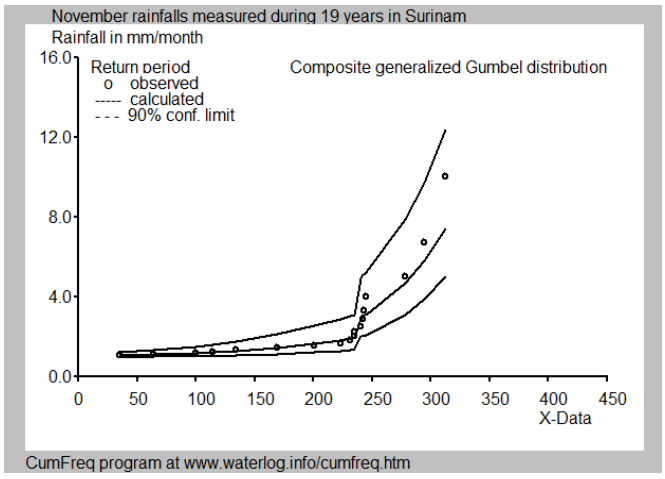


Fig.9. Return period in years for the composite generalized Gumbel distribution (Fig. 6.4).

8 Conclusions

The generalization of probability distributions enhances their applicability by establishing an improved goodness of fit. On top of that, the compartmentation realized by composite distributions helps in producing a still better fit in cases where the stochastic variable is influenced by periodically changing external conditions like shown in the examples with November rainfalls in Suriname.

The knowledge and use of generalized and composite distributions is very limited. Software for these distributions is hardly available. Therefore, in this article, provisions are made to fill the gaps.

The CumFreqA calculator package introduced here can be instrumental to apply generalized and composite probability distributions through selection options in the user interface.

References:

[1] Dr. Michael P. McLaughlin, 1993. *A compendium of Common Probability distributions*. University of Barcelona. http://www.ub.edu/stat/docencia/Diplomatura/Compendium.pdf

[2] Alamilla-López Jorge Lui, 2015. *An Approximation to the Probability Normal Distribution and its Inverse*. Ingeniería Investigación y Tecnología, volumen XVI (número 4), octubre-diciembre 2015: 605-611. ISSN 1405-7743 FI-UNAM. http://www.scielo.org.mx/pdf/iit/v16n4/v16n4a12.pdf

[3] Raminta Stockute and Paul Johnson, 2013. *Laplace distribution*. University of Kansas. http://pj.freefaculty.org/guides/stat/Distributions/DistributionWriteups/Laplace/Laplace-03.pdf

[4] Shaifar Annuar Abu Akbar et al., 2016. *GenDist, An R package for Generalized Probability Distribution Models*. In Plos Journal. https://doi.org/10.1371/journal.pone.0156537

[5] Sandra Teodorescu and Raluca Vernic, 2013. *On composite Pareto models.y Distribution Models*. Institute of Mathematical statistics and Applied Mathematics of of Romanian Academy. In Plos Journal. http://www.csm.ro/reviste/Mathematical_Reports/Pdfs/2013/1/2_Teodorescu.pdf

[6] Engineering Statistics Handbook, 1.3.6.6.9. *Lognormal Distribution*. The Information Technology Laboratory (ITL), National Institute of Standards and Technology (NIST). https://www.itl.nist.gov/div898/handbook/eda/section3/eda3669.htm

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

[8] Chapter 6: *Frequency and regression analysis.* In: H.P.Ritzema (Ed.), Drainage Principles and Applications, Publication 16, second revised edition, 1994, International Institute for Land Reclamation and Improvement (ILRI), Wageningen, The Netherlands. https://www.waterlog.info/pdf/freqtxt.pdf

[9] CumFreqA, *Free calculator for probability Distribution fitting* using generalized, mirrored, and composite distribution functions. https://www.waterlog.info/composite.htm

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

[11] *Use of Student's t-distribution* to determine confidence limits given the average and standard deviation of data in a sample. https://www.waterlog.info/t-tester.htm

[12] PennState Eberly College of Science. *ProbabilityDensity Functions*. The Pennsylvania State University. https://onlinecourses.science.psu.edu/stat414/node/97/