Statistical Modelling of Extremes with Distributions of Fréchet and Gumbel: Parameter Estimation and Demonstration of Meteorological Applications

Hristo Chervenkov, Krastina Malcheva*

Department of Meteorology
National Institute of Meteorology and Hydrology
Bulgarian Academy of Sciences
66 Tsarigradsko Shose Blvd.
1784 Sofia Bulgaria
E-mails: hristo.tchervenkov@meteo.bg, krastina.malcheva@meteo.bg

**Corresponding author*

Received: April 20, 2017 **Accepted: February 12, 2018**

Published: March 31, 2018

Abstract: *The estimation of the Cumulative Distribution Function (CDF) of data sets of random variables is a fundamental goal in the statistical modelling of extreme natural and technological accidents. Various geophysical events such as drought, floods, avalanches and heavy precipitation are a prerequisite for serious damage and economic losses, and therefore the comprehensive knowledge of their risk of occurrence, repetition periods and return levels is crucial for the planning and elaboration of mitigation strategies. The paper describes step-by-step the derivation of the parameters of the Fréchet and Gumbel CDFs, which form together with the Weibull one the Generalized Extreme Value (GEV) distribution family and are widely used for statistical modelling of extreme events. Two methods for estimation of the CDF-parameters are considered: least-square estimation and maximum-likelihood estimation. The developed and freely-available source code, written in FORTRAN 90/95, enhances the practical value of the presented work. The possibilities of the proposed approach for climatological applications are demonstrated by examples with time series of point measurements and gridded data sets.*

Keywords: *Fréchet and Gumbel distribution, CDF-parameters, Least-square estimation, Maximum-likelihood estimation, Free source code.*

Introduction

Estimating the type and parameters of the cumulative distribution function (CDF) of data set of random variables is a fundamental goal in many fields in which the analysts are interested in estimating the risk of occurrence of a particular event, for example, the probability of a catastrophic accident (drought, floods, avalanche, breakdown of technological structures, etc.).

Due to the major damage and the consequent social and economic effects, the precipitation and temperature extremes have received lots of attention from both governments and the public. Understanding of whether and how the frequency and magnitude of these extremes have changed during the past several decades is not only the focus in hydrological, meteorological, climatic, and the related studies, but also a crucial issue for the management of the associated risks. Probability distribution models are useful tools for the statistical description of heavy precipitation and extreme high/low temperatures. However, how to choose an appropriate model for a specific study is still a matter of debate [19]. Many kinds of probability distributions are available to investigate the climate extremes and, consequently, a significant number of publications are dedicated to the estimation of their appropriateness for a description of the historical frequency and the spatio-temporal variations of extreme events over the world (for a comprehensive review see [20] and the references therein). Studies show that the applicability of different probability distributions depends on the spatial and temporal differences in the considered domain. Generally, there is a consensus that only one CDF with adequate performance under all conditions is absent.

The main aim of this article is to present the applicability of two pragmatic, but at the same time mathematically consistent, approaches for estimation of parameters of two concrete probability distributions, rather then to asses the applicability of a whole group of CDFs over certain data sets. The approach is based on the two most widely used methods – the least-squares estimation (LSE) and the maximum-likelihood estimation (MLE). The straightforward procedure is very suitable for preparing a computational program as the one proposed in this study. It is written in FORTRAN 90/95 and is available to download free of charge.

Despite the importance of the issue, the existing literature appears very limited. One possible reason is the barrier-free availability of front-end solutions in form of some software packages. A similar problem, mainly from the theoretical point of view, is addressed in some articles. Thus, the paper of Abbas and Yincai [1] deals with the estimation of scale parameter for Fréchet distribution with known shape by means of MLE and probability weighted moment estimation. The same authors consider in [2] MLE and LSE of the same distribution for example of data set with lifetimes of technological items. Subject of other studies in this group are extreme meteorological events: Vivekanandan [27] presents the methodology adopted in determination of parameters of Gumbel and Fréchet distributions modelling the extreme rainfall for Fatehabad and Tohana regions in India. Six different estimation procedures, including MLE and LSE are used for determination of parameters of Gumbel CDF and one, the Order Statistics Approach, for Fréchet. As far as these studies are dedicated on more specific issues, rather then the derivation of the parameters of the CDFs, key points of the latter problem are not addressed at all. Thus, their applicability as guidance for solving real problems is limited. In contrast, the interesting and useful study of Ghosh [11] is dedicated on the determination of the parameters of the Weibull distribution and offers as well the corresponding source code. Our article closely follows the clear and punctual style of the Ghosh's one in the first section, but is focused on the Fréchet and Gumbel CDFs, providing source code for all three distributions, members of the GEV-family. The description of the step-by-step procedure, apart the availability of the source code, according the authors' opinion, is one of the strengths of the described work.

The paper is organized as follows. The description of the theoretical background is presented in the first section. The second section is dedicated on the computational procedure and its validation. Additionally, one of its possible applications is demonstrated with a data set of annual maximum daily precipitation. The third section deals with the computation of different return levels for time series of point measurements and gridded data set. Brief concluding remarks and concise comments are presented in the last section.

Theoretical background

The Weibull distribution was proposed by the Swedish mathematician Weibull [28] for description of the life length of materials under fatigue and fracture loads. This distribution is used extensively in the last decades for statistical modelling of extreme geophysical events as well. It is a member of the Generalized Extreme Value (GEV) distribution family of continuous probability distributions, including also Gumbel and Fréchet types [5].

The Weibull CDF is described as:

$$P_f(x;\, m,\, x_0,\, x_u) := 1 - \exp\left\{-\left(\frac{x - x_u}{x_0}\right)^m\right\} \text{for}\, x \geq x_u,\, x_0 > 0,\, m > 0, \tag{1}$$

where x_0, m and x_u are defined as scale parameter, shape parameter and location parameter, respectively.

If the value of m is higher, the distribution of the measured values is narrower and, subsequently, its peak is higher. Different values of the shape parameter (the Weibull modulus) can have noticeable effects on the behavior of the distribution and can complicate MLE when m is close to 1. A change in the scale parameter x_0 has the same influence on the distribution as a change of the abscissa scale. Modifying the value of x_u has the effect of sliding the distribution.

If, as a conservative approach, x_u is assumed to be zero, the resulting distribution is known as the two-parameter Weibull distribution:

$$P_f(x;\, m,\, x_0) := 1 - \exp\left\{-\left(\frac{x}{x_0}\right)^m\right\}. \tag{2}$$

The probability density function (PDF) of the two-parameter Weibull distribution is:

$$f(x;\, m,\, x_0) := \frac{\partial}{\partial x} P_f(x;\, m,\, x_0) = \frac{m}{x_0}\left(\frac{x}{x_0}\right)^{m-1} \exp\left\{-\left(\frac{x}{x_0}\right)^m\right\}. \tag{3}$$

Fréchet distribution was introduced by French mathematician Fréchet [10] in 1927 as a possible limit distribution of the largest order statistics. The Fréchet distribution has been used as a useful method for modelling and analyzing several extreme events in the nature and technology. The CDF, proposed by Fréchet, is described as:

$$P_f(x;\, m,\, x_0,\, x_u) := \exp\left\{-\left(\frac{x - x_u}{x_0}\right)^{-m}\right\} \text{for}\, x \geq x_u,\, x_0 > 0,\, m > 0, \tag{4}$$

where the meaning of the parameters x_0, m and x_u is the same as in the Weibull distribution. After the conservative approach $x_u = 0$, we get the two-parameter Fréchet distribution:

$$P_f(x;\, m,\, x_0) := \exp\left\{-\left(\frac{x}{x_0}\right)^{-m}\right\}. \tag{5}$$

The two-parameter Fréchet PDF is defined as:

$$f(x;\, m,\, x_0) := \frac{\partial}{\partial x} P_f(x;\, m,\, x_0) = \frac{m}{x_0}\left(\frac{x}{x_0}\right)^{-m-1} \exp\left\{-\left(\frac{x}{x_0}\right)^{-m}\right\}. \tag{6}$$

The Fréchet is also known as type 2 extreme value or the inverse Weibull, whereas the distribution of the negative of the Weibull random variable is a type 3 extreme value distribution. The MLE and the LSE of the parameters of the inverse Weibull distribution have been discussed by Calabria and Pulcini [4]. Implementation of the Fréchet distribution in various engineering applications have been reported in [14]. In meteorology, Zaharim et al. [30] use it for statistical modelling of the wind characteristics.

The Gumbel distribution, known also as log-Weibull distribution or the double exponential distribution, was proposed by the German mathematician Gumbel [12]. The Gumbel CDF is described as:

$$P_f(x;\, x_0,\, x_u) := \exp\left\{-\exp\left\{-\frac{x - x_u}{x_0}\right\}\right\}. \tag{7}$$

The PDF of Gumbel distribution is given as:

$$f(x;\, x_0) := \frac{\partial}{\partial x} P_f(x;\, m,\, x_0) = \frac{1}{x_0}\exp\left\{-\frac{x - x_u}{x_0} - \exp\left\{-\frac{x - x_u}{x_0}\right\}\right\}. \tag{8}$$

The absence of the scale parameter m, however, is not the main difference between the Gumbel distribution and Weibull and Fréchet distributions. As far as

$$P_f^{Weibull}(x = 0;\, x_0) = \lim_{x \to 0} P_f^{\text{Fréchet}}(x;\, x_0) = 0 \tag{9}$$

the two-parameter Weibull and Fréchet distributions are completely sufficient in the common case for description of positively definite quantities. By the Gumbel CDF however

$$P_f^{Gumbel}(x = 0;\, x_0) = \frac{1}{e} \approx 0.37 \tag{10}$$

and thus the Gumbel CDF is not applicable without location parameter.

Similarly to the previous two, the Gumbel CDF is used for many applications in the geophysical sciences, for instance, description of wind gusts [8, 24]. Klein Tank et al. [18] observes the application of this distribution for description of extreme weather events in climate change context. Palutikof et al. [24] describes and reviews methods for calculation of extreme wind speeds, including "classical" ones based on the GEV distribution. They find, that the Gumbel CDF is the most commonly used distribution applied to a set of annual maxima.

Due to the reason pointed above, the subsequent discussion in this paper will be restricted to the two- parameter Weibull and Fréchet distributions and the Gumbel's one given by Eq. (7).

The parameters of the considered distributions can be estimated by several different methods, the subject of many publications (see, for instance, [11] and [24] and references therein). Most widely used, however, are the least-square estimation of logarithmic transformed data and maximum likelihood estimation, which will be addressed in the next subsections. The derivation of the Weibull CDF-parameters is described in detail in [11] and thus the respective procedure will be skipped.

Least-square estimation

The ordinary least-squares regression on appropriately transformed data is a standard approach in regression analysis. The LSE method, applied herein, consists of linear regression with ordinary least-squares of the logarithmic transformed data. In the case of Fréchet distribution, the twofold logarithmation of both sides of Eq. (5) leads to:

$$y(x) := -\ln\left\{-\ln\left[P_f\left(x;\ m,\ x_0\right)\right]\right\} = -m\ln x_0 + m\ln x. \tag{11}$$

Similarly, for the Gumbel distribution:

$$y(x) := -\ln\left\{-\ln\left[P_f\left(x;\ x_0,\ x_u\right)\right]\right\} = \frac{x - x_u}{x_0}. \tag{12}$$

Eq. (11) and Eq. (12) are linear models: of $y(x)$ versus $\ln x$ in the first case (with a slope m and a y-intercept $-m\ln x_0$), and of $y(x)$ versus x in the second case (with a slope $1/x_0$ and a y-intercept $-x_u/x_0$).

The probability, P_f, for a given x can be calculated from n measured data after ordering in ascending order $x_1 \leq x_2 \leq \cdots \leq x_n$ and obtaining the empirical estimation. These estimates are known also as plotting positions [24]. Choosing plotting positions which lead to unbiased quantile estimates is not straightforward, and the literature is large, with at least ten published formulae [13]. Palutikof et al. [24] and Ghosh [11], however, recommend to use the following unbiased estimator (see [29] for details):

$$P_f = \frac{i}{n+1}, \tag{13}$$

where i is the rank of the ordered values of x. According to Ghosh [11], the above form gives the minimum variance.

Substituting Eq. (13) in Eq. (11) and Eq. (12), we get correspondingly:

$$y_i = -\ln\left[-\ln\left(\frac{i}{n+1}\right)\right] = -m\ln x_0 + m\ln x_i, \tag{14}$$

$$y_i = -\ln\left[-\ln\left(\frac{i}{n+1}\right)\right] = \frac{x_i - x_u}{x_0}. \tag{15}$$

In the case of the Fréchet distribution, the estimates of the slope and y-intercept using ordinary least squares are:

$$\hat{m} = \frac{n\sum\limits_{i=1}^{n} y_i \ln x_i - \sum\limits_{i=1}^{n} y_i \sum\limits_{i=1}^{n} \ln x_i}{n\sum\limits_{i=1}^{n} \ln^2 x_i - \left(\sum\limits_{i=1}^{n} \ln x_i\right)^2} \tag{16}$$

and subsequently

$$\hat{x}_0 = \exp\left\{\frac{1}{n}\left(\sum_{i=1}^{n} \ln x_i - \frac{1}{\hat{m}}\sum_{i=1}^{n} y_i\right)\right\}. \tag{17}$$

In the case of the Gumbel distribution we get:

$$\hat{x}_0 = \frac{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}{n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i} \tag{18}$$

and subsequently

$$\hat{x}_u = \frac{1}{n} \left(\sum_{i=1}^{n} x_i - \hat{x}_0 \sum_{i=1}^{n} y_i \right). \tag{19}$$

The relatively easy application of LSE determines its popularity. The method implementation as unbiased and minimum variance estimator, implicitly assumes that the error in one measurement is uncorrelated with the error in any other and the errors are normally distributed with zero mean and constant variance. In the common case, however, the satisfaction of these conditions is not guaranteed.

Maximum-likelihood estimation

The maximum-likelihood method provides a procedure for deriving the estimates of the considered distribution parameters directly. The MLE is a standard and widely adopted estimation technique that can be applied to any statistical distribution. Hence its explanation can be found in any statistical guidebook, as for instance [16], only the basics will be considered here.

Suppose that the given data set $x_1, x_2, \ldots, x_n$ is a sample of independent and identically distributed observations, coming from any distribution with a PDF with unknown parameters $a_1, a_2, \ldots, a_k$. The likelihood of obtaining a particular sample value x_i may be assumed to be proportional to the PDF at x_i. Hence, the likelihood of obtaining n independent observations $x_1, x_2, \ldots, x_n$ is:

$$f(x_1, x_2, \ldots, x_n; a_1, a_2 \ldots a_k) = f(x_1; a_1, a_2, \ldots, a_k) \cdots \times f(x_n; a_1, a_2, \ldots, a_k). \tag{20}$$

If we look at this function from a different perspective by considering the data set $x_1, x_2 \ldots x_n$ to be fixed "parameters" of this function, whereas $a_1, a_2 \ldots a_k$ will be the function's variables and allowed to vary freely; this function will be called the likelihood $\mathscr{L}$:

$$\mathscr{L}(x_1, x_2 \ldots x_n; a_1, a_2 \ldots a_k) := f(x_1, x_2 \ldots x_n; a_1, a_2 \ldots a_k) = \prod_{i=1}^{n} f(x_i; a_1, a_2 \ldots a_k) \tag{21}$$

The maximum-likelihood estimator of the parameters $a_1, a_2 \ldots a_k$ will then be the particular values of each a_i so that $\mathscr{L}$ in Eq. (21) or the probability of obtaining the data set is maximized. Due to the multiplicative nature of $\mathscr{L}$, it is generally more convenient to maximize the logarithm of the likelihood function. Taking the logarithm of the Eq. (21) breaks down the product in additions, which are then maximized separately, as follows:

$$\begin{aligned}&\frac{\partial}{\partial a_1}\ln\mathscr{L}(x_1,\, x_2,\, \ldots,\, x_n;\, a_1,\, a_2\, \ldots\, a_k) = 0\\ &\qquad\qquad \ldots\\ &\frac{\partial}{\partial a_k}\ln\mathscr{L}(x_1,\, x_2,\, \ldots,\, x_n;\, a_1,\, a_2,\, \ldots,\, a_k) = 0.\end{aligned} \tag{22}$$

The likelihood function for Fréchet distribution can be obtained by substituting Eq. (6) in Eq. (21):

$$\begin{aligned}\mathscr{L}(x_1,\, x_2,\, \ldots,\, x_n;\, m,\, x_0) = {}&\frac{m^n}{x_0^n}\left[\left(\frac{x_1}{x_0}\right)^{-1-m}\left(\frac{x_2}{x_0}\right)^{-1-m}\ldots\left(\frac{x_n}{x_0}\right)^{-1-m}\right]\times\\ &\exp\left\{-\left(\frac{x_1}{x_0}\right)^{-m}\right\}\exp\left\{-\left(\frac{x_2}{x_0}\right)^{-m}\right\}\ldots\\ &\exp\left\{-\left(\frac{x_n}{x_0}\right)^{-m}\right\} =\\ &\frac{m^n}{x_0^n}\prod_{i=1}^{n}\left(\frac{x_i}{x_0}\right)^{-1-m}\times\prod_{i=1}^{n}\exp\left\{-\left(\frac{x_i}{x_0}\right)^{-m}\right\}.\end{aligned} \tag{23}$$

Taking the logarithm of the both sides and rearranging the terms yields:

$$\ln\mathscr{L}(x_1,\, x_2,\, \ldots,\, x_n;\, m,\, x_0) = n\ln m - n\ln x_0 - (m+1)\sum_{i=1}^{n}\ln\frac{x_i}{x_0} - \sum_{i=1}^{n}\left(\frac{x_i}{x_0}\right)^{-m}. \tag{24}$$

Taking the derivative of $\ln\mathscr{L}(x_1,\, x_2,\, \ldots,\, x_n;\, m,\, x_0)$ with respect to x_0 equating it to 0, we get:

$$x_0^m = \frac{n}{\sum\limits_{i=1}^{n} x_i^{-m}}. \tag{25}$$

Similarly, equating the derivative of $\ln\mathscr{L}(x_1,\, x_2,\, \ldots,\, x_n;\, m,\, x_0)$ with respect to m to 0 yields:

$$\frac{n}{m} + n\ln x_0 - \sum_{i=1}^{n}\ln x_i + \sum_{i=1}^{n}\left(\frac{x_i}{x_0}\right)^{-m}\ln\frac{x_i}{x_0} = 0. \tag{26}$$

Substituting Eq. (25) in Eq. (26), we get finally:

$$\frac{n}{m} - \sum_{i=1}^{n}\ln x_i + \frac{n\sum\limits_{i=1}^{n} x_i^{-m}\ln x_i}{\sum\limits_{i=1}^{n} x_i^{-m}} = 0. \tag{27}$$

The unknown parameters x_0 and x_u of the Gumbel distribution can be found in a similar way.

Thus, after substituting Eq. (8) in Eq. (21), the likelihood function can be written as:

$$\begin{aligned}\mathscr{L}(x_1,\, x_2,\, \ldots,\, x_n;\, x_0,\, x_u) = \exp\left\{-\frac{x_1-x_u}{x_0} - \exp\left\{-\frac{x_1-x_u}{x_0}\right\}\right\}\times \\ \exp\left\{-\frac{x_2-x_u}{x_0} - \exp\left\{-\frac{x_2-x_u}{x_0}\right\}\right\}\times \\ \cdots \\ \exp\left\{-\frac{x_n-x_u}{x_0} - \exp\left\{-\frac{x_n-x_u}{x_0}\right\}\right\} = \\ \prod_{i=1}^{n}\exp\left\{-\frac{x_i-x_u}{x_0} - \exp\left\{-\frac{x_i-x_u}{x_0}\right\}\right\}.\end{aligned} \tag{28}$$

Taking the logarithm of both sides and accounting that $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$,

$$\ln\mathscr{L}(x_1,\, x_2,\, \ldots,\, x_n;\, x_0,\, x_u) = -n\ln x_0 - \frac{1}{x_0}\sum_{i=1}^{n}(x_i - x_u) - \sum_{i=1}^{n}\exp\left\{-\frac{x_i-x_u}{x_0}\right\}. \tag{29}$$

Taking the derivative of $\ln\mathscr{L}(x_1,\, x_2\, \ldots\, x_n;\, x_0,\, x_u)$ with respect to x_0 equating it to 0, we get:

$$-\frac{n}{x_0} + \frac{1}{x_0^2}\sum_{i=1}^{n}(x_i - x_u) - \frac{1}{x_0^2}\sum_{i=1}^{n}(x_i - x_u)\exp\left\{-\frac{x_i-x_u}{x_0}\right\} = 0. \tag{30}$$

Similarly, equating the derivative of $\ln\mathscr{L}(x_1,x_2,\ldots,\, x_n;x_0,x_u)$ with respect to x_u to 0 yields:

$$\frac{n}{x_0} - \frac{1}{x_0}\sum_{i=1}^{n}\exp\left\{-\frac{x_i-x_u}{x_0}\right\} = 0, \tag{31}$$

which can be rewritten as:

$$x_u = x_0 \ln\frac{n}{\sum_{i=1}^{n}\exp\left\{-\frac{x_i}{x_0}\right\}}. \tag{32}$$

Substituting Eq. (32) in Eq. (30), after rearranging the terms, we get the expression for x_0:

$$\bar{x} - x_0 - \frac{\sum_{i=1}^{n} x_i \exp\left\{-\frac{x_i}{x_0}\right\}}{\sum_{i=1}^{n}\exp\left\{-\frac{x_i}{x_0}\right\}} = 0. \tag{33}$$

Eq. (27) and Eq. (33) are equations with respect to m and x_0 correspondingly which roots can be found numerically with different methods (see, for instance, [26]). According to the Newton-Raphson iterative method, for example, the recurrence relation between the i^{th} and $(i+1)^{th}$

approximation of the root of the equation is:

$$z_{k+1} = z_k - \frac{\phi'(z)}{\phi(z)}, \tag{34}$$

where $\phi(z)$ is the function of the left-hand side of Eq. (27) or Eq. (33) and z is m or x_0. Due to two reasons, this method is appropriate for the considered problem. First, it is very powerful: it converges quadratically. The number of significant digits approximately doubles at each step near the root [25]. Second, and most important, the derivative $\phi'(z)$ can be obtained analytically. Thus, for the Fréchet distribution we have:

$$\phi'(m) = -\frac{n}{m^2} + n\frac{\partial}{\partial m}\frac{n\sum\limits_{i=1}^{n} x_i^{-m}\ln x_i}{\sum\limits_{i=1}^{n} x_i^{-m}} \tag{35}$$

or

$$\phi'(m) = -\frac{n}{m^2} + n\frac{\left(\sum\limits_{i=1}^{n} x_i^{-m}\ln x_i\right)^2 - \sum\limits_{i=1}^{n} x_i^{-m}\ln^2 x_i \sum\limits_{i=1}^{n} x_i^{-m}}{\left(\sum\limits_{i=1}^{n} x_i^{-m}\right)^2}. \tag{36}$$

Substituting Eq. (36) in Eq. (34) we get finally:

$$m_{k+1} = m_k - \frac{\dfrac{1}{m_k} - \dfrac{\sum\limits_{i=1}^{n}\ln x_i}{n} + \dfrac{\sum\limits_{i=1}^{n} x_i^{-m_k}\ln x_i}{\sum\limits_{i=1}^{n} x_i^{-m_k}}}{-\dfrac{1}{m_k^2} + \dfrac{\left(\sum\limits_{i=1}^{n} x_i^{-m_k}\ln x_i\right)^2 - \sum\limits_{i=1}^{n} x_i^{-m_k}\ln^2 x_i \sum\limits_{i=1}^{n} x_i^{-m_k}}{\left(\sum\limits_{i=1}^{n} x_i^{-m_k}\right)^2}}. \tag{37}$$

Similarly, for the case of the Gumbel distribution:

$$\phi'(x_0) = -1 - \frac{\partial}{\partial x_0}\frac{\sum\limits_{i=1}^{n} x_i \exp\left\{-\dfrac{x_i}{x_0}\right\}}{\sum\limits_{i=1}^{n} \exp\left\{-\dfrac{x_i}{x_0}\right\}} \tag{38}$$

or

$$\phi'(x_0) = -1 - \frac{\sum\limits_{i=1}^{n} x_i^2 \exp\left\{-\dfrac{x_i}{x_0}\right\}\sum\limits_{i=1}^{n} \exp\left\{-\dfrac{x_i}{x_0}\right\} - \left(\sum\limits_{i=1}^{n} x_i \exp\left\{-\dfrac{x_i}{x_0}\right\}\right)^2}{x_0^2\left(\sum\limits_{i=1}^{n} \exp\left\{-\dfrac{x_i}{x_0}\right\}\right)^2}. \tag{39}$$

Substituting Eq. (39) in Eq. (34) we get finally:

$$x_{0\,k+1} = x_{0\,k} + \frac{\bar{x} - x_{0\,k} - \dfrac{\sum\limits_{i=1}^{n} x_i \exp\left\{-\dfrac{x_i}{x_{0\,k}}\right\}}{\sum\limits_{i=1}^{n} \exp\left\{-\dfrac{x_i}{x_{0\,k}}\right\}}}{1 + \dfrac{\sum\limits_{i=1}^{n} x_i^2 \exp\left\{-\dfrac{x_i}{x_{0\,k}}\right\} \sum\limits_{i=1}^{n} \exp\left\{-\dfrac{x_i}{x_{0\,k}}\right\} - \left(\sum\limits_{i=1}^{n} x_i \exp\left\{-\dfrac{x_i}{x_{0\,k}}\right\}\right)^2}{x_{0\,k}^2 \left(\sum\limits_{i=1}^{n} \exp\left\{-\dfrac{x_i}{x_{0\,k}}\right\}\right)^2}}. \tag{40}$$

From an initial guess of m in Eq. (37) and x_0 in Eq. (40), the value of Fréchet modulus and the Gumbel scale parameter can be estimated when the difference between subsequent iterations is less than a predefined tolerance value.

Small parts for the described above procedure can be found in some reports. Thus, for example, Eq. (3) in the [23] is equivalent of Eq. (29), but explicit expressions for m as in Eq. (37) and x_0 as in Eq. (40), which are the final outcome and are most important in pragmatic sense, are missing. Under the many strengths of the MLE can be emphasized its consistency and the fact, that the method, in contrast to the LSE, does not rely on any empirical functions as Eq. (13). A certain drawback is, at least from a practical point of view, the complicated procedure of estimation of the CDF-parameters.

Program implementation, validation and demonstration

The computer program has been written in FORTRAN 90/95 and is available upon request. It consists of three identical sections, one for each distribution (Weibull, Fréchet and Gumbel). The first one uses some re-syntaxed segments from the Ghosh's program WEIBUL [11], preserving part of their functionality. The current implementation, however, is significantly improved, using, in particular, the possibilities of the modern language. The program computes the parameters m and x_0 of both Weibull and Fréchet CDFs, as well as x_0 and x_u for the Gumbel distribution, according to the described above LSE and MLE methods. The concordance of the obtained continuous distributions with the empirical one is estimated with the root-mean-square error (RMSE).

The parameter calculations were validated by comparing results, concerning the Weibull distribution, with the outcome of Ghosh's program, using the data set of Ang and Tang [3]. The outcome of both programs is the same. Although the maximum number of iterations is set to 50, in all performed tests the method converges to the solution using an initial value of m and x_0 equal to 1 and tolerance 1×10^{-4} in less than 10 iterations.

Although relatively small, the climate of Bulgaria is diverse. Its territory is divided in two climate areas – European-Continental climate area and Continental-Mediterranean one and four subareas (Moderate-Continental, Transition-Continental, South-Bulgarian and Black-Sea one) [21]. The program possibilities are demonstrated with two examples, handling data sets from two stations of the meteorological network of the National Institute of Meteorology and Hydrology – Bulgarian Academy of Sciences (NIMH–BAS). The stations are purposefully selected with very different climate conditions. The series of annual maximum daily precipitation will be considered as "precipitation extremes" herein.

The first station is Vratsa. Its coordinates are 43.23N; 23.53E and is located on the foot of the Balkan mountain, belonging to the European-Continental climate area. The station is in operational use since May 1929 and the absolute maximum of precipitation is 109.3 mm recorded on 22.08.1966. The second station is Resovo. Its coordinates are 41.99N; 28.03E and it is the southernmost Bulgarian coastal station, belonging to the Black-Sea climate subarea. Data in the period 1961-2015 are used, the absolute maximum of precipitation is 216.7 mm recorded on 25.09.1977. This value is extremely high for the Bulgarian climate conditions. Figs. 1 and 2 depict the observed values and the fitted CDFs, calculated with the parameters from the program output.

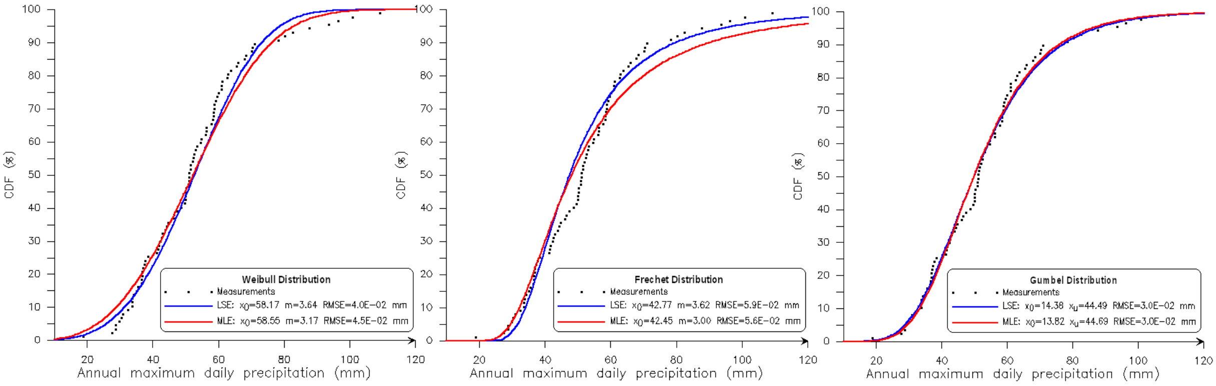


Fig. 1 Measured precipitation extremes and fitted Weibull (left pane), Fréchet (middle pane) and Gumbel (right pane) CDFs for station Vratsa

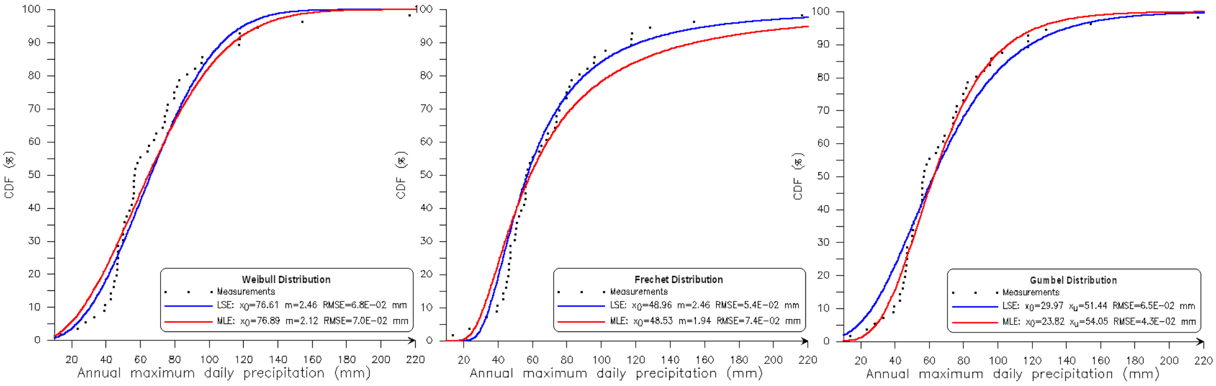


Fig. 2 Measured precipitation extremes and fitted Weibull (left pane), Fréchet (middle pane) and Gumbel (right pane) CDFs for station Resovo

Return levels estimation

The estimation of the return levels (RL) of extreme events on the basis of a much shorter period of observations, i.e. what extreme values might occur in 20-, 50-, 100-year or even longer, is very important task [20]. The T-year, where T is the period of occurance, RL can be calculated straightforwardly once the model parameters are estimated. For example, let $\hat{m}$ and $\hat{x}_0$ be the Weibull CDF-parameters, obtained with LSE or MLE, then, according to Eq. (2), we get:

$$x = \hat{x}_0 \left[\ln \frac{1}{P_f\left(x;\, m,\, x_0\right)} \right]^{\frac{1}{\hat{m}}} \tag{41}$$

and substituting $P_f = 1 - 1/T$:

$$x = \hat{x}_0 \left(\ln T\right)^{\frac{1}{\hat{m}}}. \qquad (42)$$

Similarly, from the Fréchet distribution we get:

$$x = \hat{x}_0 \left[\ln\left(\frac{T}{T-1}\right)\right]^{-\dfrac{1}{\hat{m}}} \qquad (43)$$

and for the Gumbel distribution:

$$x = \hat{x}_u - \hat{x}_0 \ln\left[\ln\left(\frac{T}{T-1}\right)\right] \qquad (44)$$

in which T means the return period and x denotes the theoretical RL for a given return period.

The 100-, 50-, 20- and 5-year RLs for seek of brevity for the Vratsa data set only, estimated with the obtained parameters, are listed in Table 1.

Table 1. Estimated return levels for the Vratsa data set (in mm)

Return period, years	**Weibull CDF**		**Fréchet CDF**		**Gumbel CDF**	
	LSE	MLE	LSE	MLE	LSE	MLE
100	88.5	94.8	152.5	196.2	110.7	108.3
50	84.6	90.0	125.7	155.5	100.6	98.6
20	78.6	82.8	97.2	114.1	87.2	85.7
5	66.3	68.0	64.7	69.9	66.1	65.4

It can be seen that the largest discrepancy between the estimations occurs for the longest return period, calculated with the Weibull and Fréchet CDF, which is supported by the results, shown in Figs. 1 and 2. Both curves for the Weibull CDF overestimate the probability of no exceedance of upper extreme values resulting in slowly increasing of RL-values in this interval. The opposite is the case of the Fréchet CDF: the both curves, especially the MLE one, underestimate the empirical distribution, rising significantly even for the highest extreme values. As far as the both curves (i.e., obtained with LSE and MSE) for the both stations, appears closest to the empirical distribution for the high values, the Gumbel CDF appears most adequate for the approximation of the upper limit of the empirical distribution. In the case of the 5-year return period (80% probability of no exceedance), all CDF curves are seemingly closer to the empirical data set, and the obtained RL values are similar.

The freely available for the research community database E-OBS version 13.0 of the European Climate Assessment & data set (ECA&D) project [15] is used as a source for the computing of the extreme annual daily maximum temperature (EADMT), considered in the second example. The data set is in a regular grid of 0.25°×0.25° for the current implementation and daily temporal resolution, which is pre-processed, obtaining the annual extremes. The data-base is updated periodically, data for the period 1950-2014 inclusive are used in the second example.

The main advantage of this data set in comparison with the point observations is that all data are passed a priori quality/homogeneity control and they are represented in spatially and temporally continuous form of gridded digital map. This makes it a potentially useful source of information for monitoring long-term changes in extremes. The 50-year return level of EADMT is computed with both LSE and MLE methods, only for the grid cells with full 65-years length of the time-series, as shown on Fig. 3.

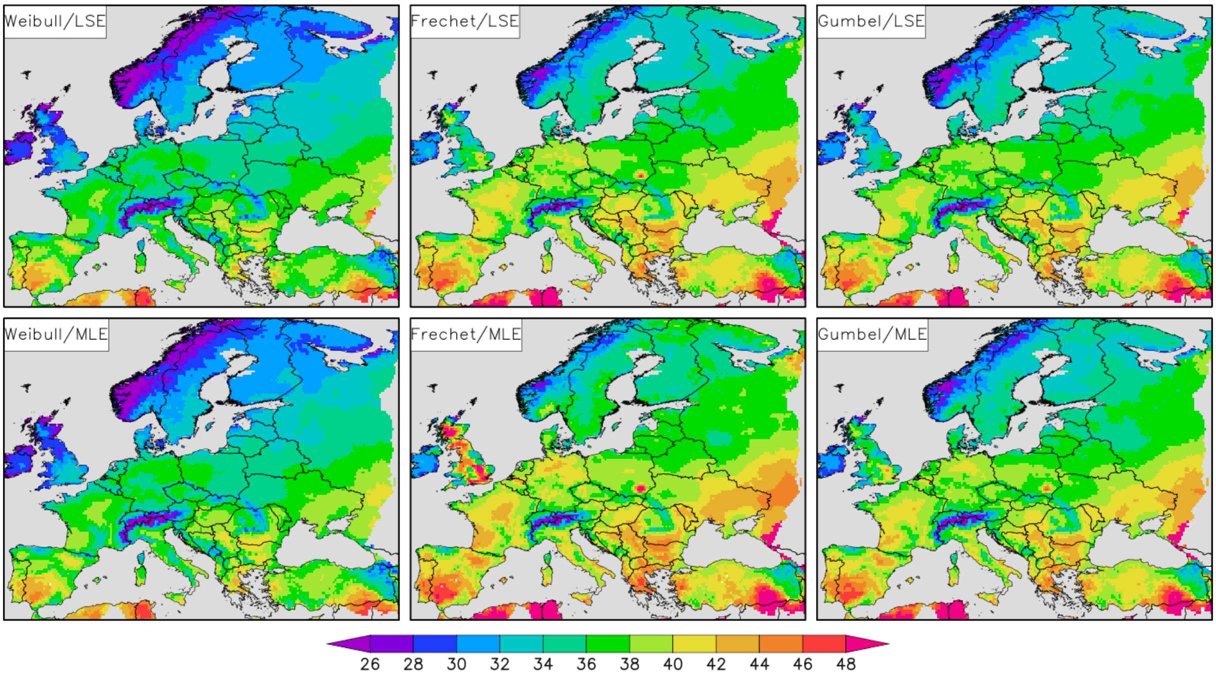


Fig. 3 50-year return level of the extreme annual daily maximum temperature (°C), obtained with the CDFs and estimators according the subplot titles (database E-OBS version 13.0)

The 50-year return level of the extreme annual daily maximum temperature is very important in planning for e.g. new infrastructure and thus it is a part of the EU-Mandated Harmonised Standards Eurocode (see https://law.resource.org/pub/eu/eurocode.html for details). Roughly speaking, the return level is similarly distributed of all panes of Fig. 3. The north-south temperature gradient and elevation effects are well reproduced. Generally, as for the case with the extreme annual precipitation, the Weibull CDF tends to produce smaller, and the Fréchet CDF to produce larger values for the 50-year RL. Again, as in the above case, the Gumbel distribution shows intermediate results. This fact is most obviously expressed over the Scandinavian mountains.

The simulated RL over a part of England on some subplots is not realistic and thus have to be detailed with example for concrete grid-cell.

The analysis reveals that this is caused by values of the EADMT in 2013, which can be quantified as outliers. This conclusion is illustrated with concrete example for two adjacent grid-cells, centered on longitude -0.875° and latitude 51.125° and 51.375° correspondingly, as shown on on Fig. 4. The EADMT values in the both grid-cells are close to each other in all years, as expected, except 2013. Astonishingly, the EADMT difference in 2013 is almost 6°C! The extremely heterogeneous distribution of the EADMT over the British islands, suggests problems in the data set preparation for this year. The replacement of the suspicious EADMTs for 2013 with these, say, from the previous year, smooths the return levels field to the background values of about 36-40 °C. Same is the case with the hot-spot in SE Poland. It is worthy to emphasize also, that these irregularities are at strongest expressed by the Fréchet CDF, and especially when its parameters are obtained with MLE. The reason could be rooted in the limit behavior of this distribution: It is heavy-tailed distributions (as, for example, these of Pareto, Student t, Cauchy, Burr, log-gamma). In contrast, the Weibul, similarly to the the uniform and beta, is short-tailed distribution with a finite right end-point. Interim case is the Gumbel CDF – like the normal, exponential, gamma and log-normal, it tails decay essentially exponentially. More details about this issue can be find in [6] and the references therein.

The analysis, performed in ECA&D (Else van den Besselaar, personal communication) reveals the reasons for the described error – a swap in the data headers for the daily maximum and the daily minimum temperature. This error is fixed in the subsequent versions 13.1 and 14.0.

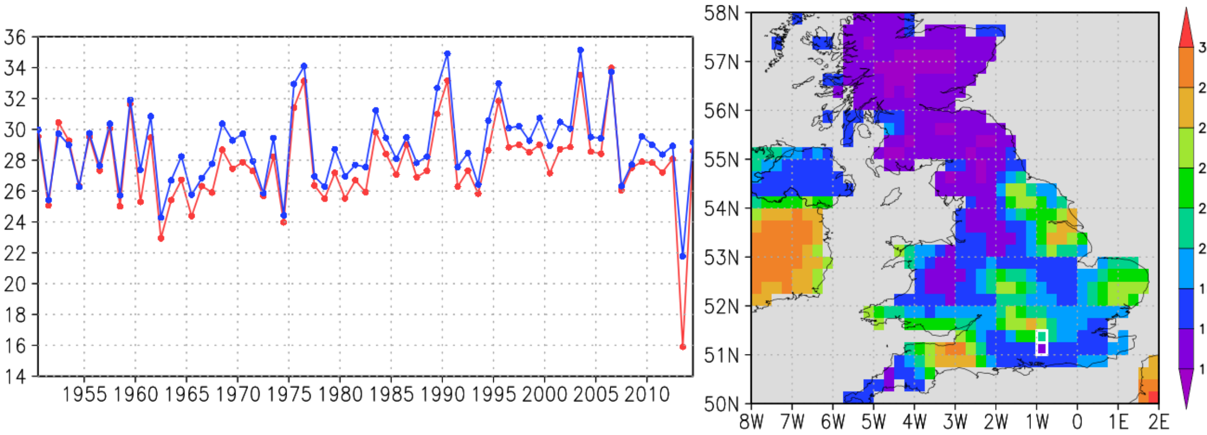


Fig. 4 Time series of the EADMT (°C) in the considered two grid-cells (left panel) and the distribution of the same variable in 2013 over part of the British islands (right panel). The frames of the grid-cells are highlighted.

In order to obtain the 50-year RL from outlier-free gridded data set, recalculation for the same period with the newest version (i.e., 16.0) of E-OBS is performed. The resulting RL-distribution is shown on Fig. 5. Apparently, the spotted above problems are not presented here with a small exception over SE Britain. Thus, the overall impression from the last two figures is the confirmed applicability of the gridded E-OBS data set for computation of return levels of temperature extremes in pan-European context.

The described shortcoming in E-OBS v13.0 shows clearly the high sensitivity of the MLE to the presence of outliers in the input data, which is higher than the LSE sensitivity (notably in the case of the Fréchet CDF).

It is well-established in the statistical literature, that the MLE does not provide consistent as well as efficient estimations for the parameters of the considered CDF in the presence of outliers (see [17] and references therein). To overcome this problem, some authors suggest modification and/or generalization of the MLE. Thus, for instance, Neykov et al. [22] propose Trimmed Likelihood Estimator (TLE) as a useful alternative to the MLE. The basic idea behind the trimming in this estimator is the removal of those observations, which values would be highly unlikely to occur. Under the weighted maximum likelihood (WML) approach, proposed by Ejaz et al. [7] and used from Khokan et al. [17] for the Weibull CDF, the weighted likelihood function $\mathscr{L}^*$ is

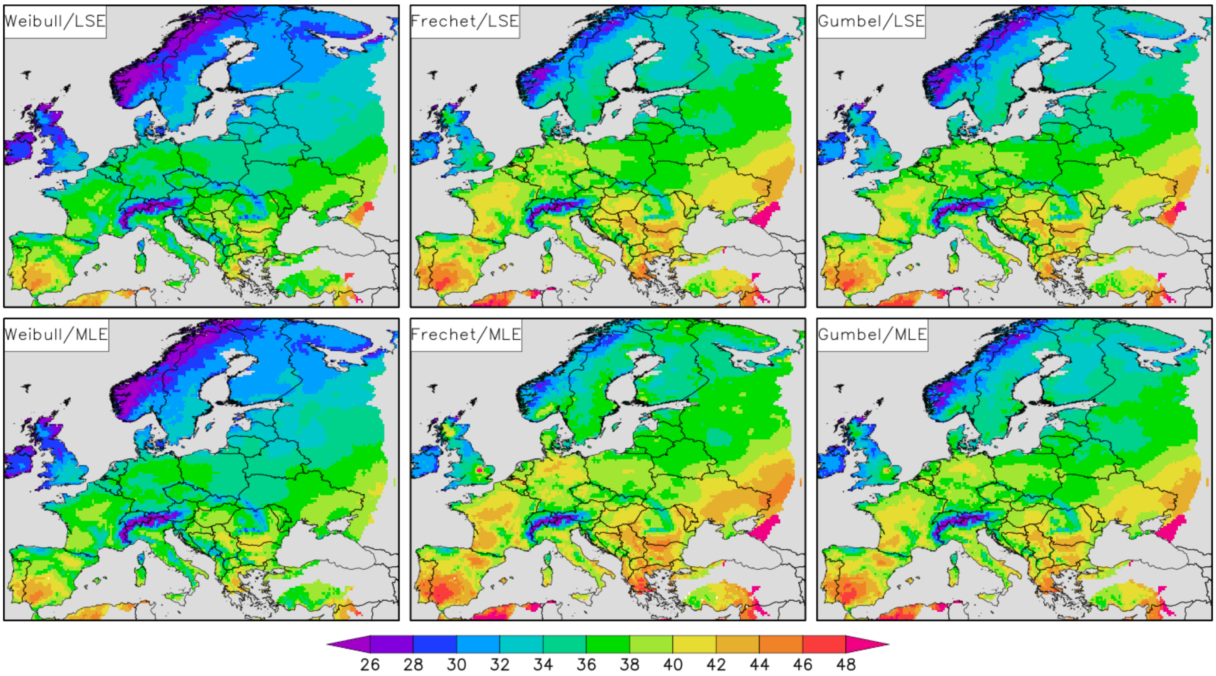


Fig. 5 50-year return level of the extreme annual daily maximum temperature (°C), obtained with the CDFs and estimators according the subplot titles (database E-OBS version 16.0)

defined as:

$$\mathscr{L}^*(x_1, x_2, \ldots, x_n; a_1, a_2, \ldots, a_k) := \prod_{i=1}^{n} f^{\delta_i}(x_i; a_1, a_2, \ldots, a_k), \tag{45}$$

where δ_i takes value 1, if the observation is not an outlying observation, otherwise it takes value 0. According to the authors, these methods show overall better performance, dealing with noisy data. The selection, however, of some procedure-specific tuning parameters supposes a degree of arbitrariness, and, generally, they are much more sophisticated and respectively computationally demanding.

Finally, the 50-year RL, calculated at NIMH – BAS for Bulgaria with data from 125 stations [9], reveal values of about 40 to 45°C for the lowlands, which are very close to those seen on Fig. 3 and Fig. 4.

Conclusion

This paper describes the estimation procedure of the Fréchet and Gumbel cumulative distribution function of data sets of random variables using the widely used methods LSE and MLE. The described explicit approach allows their effective derivation with prescribed in advance tolerance. Both methods form the basis of the developed FORTRAN 90/95 code which can be used 'as it is' or as part of other projects. The proposed solutions could be used by researchers to develop their own code, in general case, of preferable language. This was the main goal of the authors. As far as the paper primarily targets the geophysical community, two relevant examples are shown here. The high sensitivity of the MLE to the presence of outliers is also discussed. The author's opinion is that the MLE should be used carefully when dealing with noisy data sets. Although, generally, such statistical fitting can be performed with stand-alone statistical packages, in many cases the explicit code is preferable. In all cases, however, it is researcher's responsibility, to check the data set for outliers and to select the proper CDF. There is no general recipe, but well-elaborated goodness-of-fit tests should be used to estimate the feasibility of the theoretical continuous distribution.

Acknowledgments

We acknowledge the E-OBS data set from the ENSEMBLES project of European Commission's 6th Framework Programme (http://ensembles-eu.metoffice.com) and the data providers in the ECA& D project (http://www.ecad.eu).

Personal thanks to Else van den Besselaar from ECA& D for the cooperation.

References

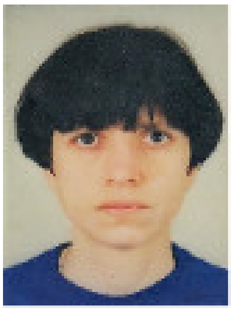
1. Abbas K., T. Yincai (2012). Comparison of Estimation Methods for Fréchet Distribution with Known Shape, Caspian Journal of Applied Sciences Research, 1(10), 58-64.
2. Abbas K., T. Yincai (2013). Estimation of Parameters for Fréchet Distribution Based on Type-II Censored Samples, Caspian Journal of Applied Sciences Research, 2(7), 36-43.
3. Ang A. H.-S., W. H. Tang (Eds.) (1984). Probability Concepts in Engineering Planning and Design, Vol. 2: Decision, Risk, and Reliability, John Wiley and Sons, New York.
4. Calabria R., G. Pulcini (1990). On the Maximum Likelihood and Least Squares Estimation in the Inverse Weibull Distribution, Statistica Applicata, 2, 53-66.
5. Coles S. G. (2001). An Introduction to Statistical Modeling of Extreme Values, Springer, Berlin, Germany.
6. Dell'Aquila R., P. Embrechts (2006). Extremes and Robustness: A Contradiction?, Financial Markets and Portfolio Management, 20(1), 103-118, https://link.springer.com/article/10.1007/s11408-006-0002-x .
7. Ejaz S. A., I. V, Andrei, A. H. Abdulkadir (2005). Robust Weighted Likelihood Estimation of Exponential Parameters, IEEE Transactions on Reliability, 54(3), 389-395.
8. Escalante-Sandoval C. A. (2013). Estimation of Extreme Wind Speeds by Using Mixed Distributions, Ingeniería Investigación y Tecnología, XIV(2), 153-162.
9. Eurocode 1 – Actions on Structures – Part 1-5: General Actions – Thermal Actions – National annex to BDS EN 1991-1-5:2005, available at http://www.bds-bg.org/images/upload/Nacionalni_prilojenia/BDS_EN_1991-1-5-NA.pdf
10. Fréchet M. (1927). Sur la loi de probabilite de lecart maximum, Annales de la Société Polonaise de Mathématique, 6, 93-116.
11. Ghosh A. (1999). A FORTRAN Program for Fitting Weibull Distribution and Generating Samples, Computers and Geosciences, 25(7), 729-738.
12. Gumbel E. J. (1958). Statistics of Extremes, Columbia University Press, New York, USA.
13. Guo S. L. (1990). A Discussion of Unbiased Plotting Positions for the General Extreme Value Distribution, Journal of Hydrology, 121(1-4), 33-44.
14. Harlow D. G. (2002). Applications of the Fréchet Distribution Function, International Journal of Material and Product Technology, 5(17), 482-495.
15. Haylock M. R., N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P. D. Jones, M. New (2008). A European Daily High-resolution Gridded Dataset of Surface Temperature and Precipitation for 1950-2006, Journal of Geophysical Research: Atmospheres, 113, D20119, doi:10.1029/2008JD10201.
16. Hazewinkel M. (Ed.) (2001). Maximum-likelihood Method, Encyclopedia of Mathematics, Springer.
17. Khokan M. R., W. Bari, J. A. Khan (2013). Weighted Maximum Likelihood Approach for Robust Estimation: Weibull Model, Dhaka University Journal of Science, 61(2), 153-156.

18. Klein Tank A., F. Zwiers, X. Zhang (2009). Guidelines on Analysis of Extremes in a Changing Climate in Support of Informed Decisions for Adaptation, Climate Data and Monitoring WCDMP-No. 72, World Meteorological Organization TD No. 1500.
19. Li Z., F. Brissette, J. Chen (2013). Finding the Most Appropriate Precipitation Probability Distribution for Stochastic Weather Generation and Hydrological Modelling in Nordic Watersheds, Hydrological Processes, 27(25), 3718-3729.
20. Li Z., Z. Li, W. Zhao, Y. Wang (2015). Probability Modeling of Precipitation Extremes over Two River Basins in Northwest of China, Advances in Meteorology, 2015, Article ID 374127, 1-13, http://dx.doi.org/10.1155/2015/374127.
21. Malcheva K., H. Chervenkov, T. Marinova (2016). Winter Severity Assessment on the Basis of Measured and Reanalysis Data, Proceedings of the 16th International Multidisciplinary Scientific Geoconference SGEM 2016, Albena, Bulgaria, 28 June – 7 July, Book 4, Vol. 1, 719-726.
22. Neykov N., R. Dimova, P. Neytchev (2005). Trimmed Likelihood Estimation of the Parameters of the Generalized Extreme Value Distributions: A Monte-Carlo Study, Pliska Studia Mathematica Bulgaria, 17, 187-200.
23. Pal K., S. Kageyama, S. Pal (2011). A Comparison of Different Procedures Relating to Parameter Estimation in Extreme Value Type I Distribution Through Simulation, Bulletin of Hiroshima Institute of Technology, Vol. 45, 281-290.
24. Palutikof J. P., B. B. Brabson, D. H. Lister, S. T. Adcock (1999). A Review of Methods to Calculate Extreme Wind Speeds, Meteorological Applications, 6(2), 119-132.
25. Press W. H., B. P. Flannery, S. A. Teukolsky, W. T. Vetterling (1986). Numerical Recipes: The Art of Scientific Computing, 355-361.
26. Stoer J., R. Bulirsch (2002). Introduction to Numerical Analysis, Texts in Applied Mathematics, 3rd Ed., Springer, New York, Vol. 12, 289-363.
27. Vivekanandan N. (2013). Comparison of Parameter Estimation Procedures of Gumbel and Fréchet Distributions for Modelling Annual Maximum Rainfall, Wyno Journal of Engineering & Technology Research, 1(1), 1-9.
28. Weibull W. (1951). A Statistical Distribution Function of Wide Applicability, Journal of Applied Mechanics, 18, 293-297.
29. Wilks S. S. (1948). Order Statistics, Bulletin of American Mathematical Society, 54, 6-50.
30. Zaharim A., S. K. Najidi, A. M. Razali, K. Sopian (2009). Analyzing Malaysian Wind Speed Data Using Statistical Distribution, Proceedings of the 4th IASME/WSEAS International Conference on Energy and Environment, University of Cambridge, 363-370.

Assoc. Prof. Hristo Chervenkov, Ph.D.
Email: hristo.tchervenkov@meteo.bg

Hristo Chervenkov received his M.Sc. degree from Sofia University "St. Kliment Ohridski", Bulgaria in 1997 and his Ph.D. in 2007 with thesis about the numerical simulation of the trans-boundary air pollution over SE Europe. He works as a senior scientist in the Department of Meteorology at the National Institute of Meteorology and Hydrology – Bulgarian Academy of Sciences (NIMH – BAS) and has many international specializations. His research interests are climatology, numerical modelling, air pollution and planetary boundary layer dynamics. He has been involved in international projects and has published more than thirty papers in reviewed journals and proceedings of international conferences.

Assist. Prof. Krastina Malcheva, M.Eng.
Email: krastina.malcheva@meteo.bg



Krastina Malcheva has M.Eng. degree in Automation and Microprocessor Systems and works as Assistant Professor in the Division of Climatology at the Department of Meteorology of NIMH – BAS since 2010. Her research interests are climatology, climate modelling and analysis of extremes. She was involved in significant national projects and has provided expert assessments for the needs of governmental organizations and institutions. She is co-author in more than twenty papers in Bulgarian Journal of Meteorology and Hydrology (http://meteorology.meteo.bg/global-change/index.html) and in proceedings of international conferences.