Revue d'Economie et de Statistique Appliquée
Volume 18 number 1, June 2021
ISSN : 1112-234X EISSN : 2600-6642
Pages : 45- 55

MODELING MAXIMUM OF MONTHLY AVERAGE TEMPERATURE USING THE GENERALIZED PARETO DISTRIBUTION AT THE JIJEL WEATHER STATION FROM 1988 TO 2018

[*]Hassen CHERAITIA

Department of Mathematics, University of Mohamed Seddik Ben Yahia, Jijel, Algeria
hassen.stat@hotmail.fr

Received: 02/01/2021 **Accepted** 22/03/202 **Published online:** 29/06/2021

ABSTRACT: The objective of this article is to estimate the two parameters of the generalized Pareto distribution (GPD) that allows us to calculate return levels of temperature for different return periods. we worked on a monthly average temperatures series of the Jijel weather station (Northern Algeria), after the determination of the threshold (u_0 = 18 °C), we estimated the parameters of the GPD by the maximum likelihood method, it was found that the generalized Pareto distribution of type II (with bounded tails) is more suitable for the maximum monthly temperatures. Once the estimated model has been validated, we calculated returns levels for different periods, according to this model. We must wait about 100 years to record an average monthly temperature of 28.56 °C.

Key Words: generalized Pareto distribution; maximum likelihood estimate, average monthly temperature; return level.

JEL Classification: C1, C13, C46, C490,

1. INTRODUCTION

The extreme temperature events can cause of various massive and destructive problems worldwide, these include cases of hospitalization, loss of lives and economic challenges. It has significant effects on health and power outages, on agriculture, such as drought. All of these effects would lead to the economic loss. One of the fundamental problems encountered in climatology is the need to establish an assessment of climate risks resulting from extreme temperature to avoid human and material damage, and therefore to provide for the occurrence of disasters and unforeseen events and if possible their intensity. The modern theory of extreme values developed between 1920 and1940 (Fréchet, 1927; Fisher and Tippet, 1928; Gnedenko, 1943; and Gumbel, 1958) finds application in many fields: finance, assurance, hydrology, meteorology...

One of the possible methods for modelling extreme temperature is the peak over threshold method (POT) (J. Pickand, 1975; Coles and Tawn, 1994; Davison and Smith, 1990; Smith, 1987; Embrechets et *al*, 1997; Reiss and Thomas, 200. The main problem is to choose the threshold and the proposed methods are not easy to implement. Under certain general conditions, the distribution of the sample of excesses above a high threshold can only belong to one of the three following laws: the Pareto type II (with bounded support), the exponential distribution (fine tails) and Pareto type I distribution (with heavy tails). Different methods can be used to estimate the parameters of the extreme laws such as the method of maximum

[*] Corresponding Authors

likelihood (Coles, 2001), the method of weighted moments (Hosking, 1990) and the Bayesian method (Smith and Naylor, 1987).

The aim of this paper is to study of monthly maximum temperature by applying the generalized Pareto distribution (GPD)[†] to the jijel weather station data in order to understand the behaviour of maximum rainfall and to establish an adequate forecasting model that helps meteorologists, insurers and authorities to understand these exceptional events and thus prevent climate risks. The zyp (Bronaugh and Werner, 2013), evd (Stephenson, 2002), extRemes (Gilleland and Katz, 2005) and ismev (Stephenson, 2014) packages of R (R core Team, 2015) were used for the data analysis. The paper is organized in the following manner, additional to this introduction: the GPD distributions, the maximum likelihood estimates of its parameters and the return level are presented in Section 2; then the theoretical model is applied to data in Section 3; finally, some conclusions are given in Section 4.

2. METHODOLOGY

The temperature data used in this paper correspond to a 31 years (from 1988 to 2018) of average, monthly temperature measured at the jijel weather station, northern Algeria. To analyse the extremes values of temperature statistically, POT method, where monthly values above a pre-determined threshold value were modelled by GPD. The GPD and Estimation method of its parameters are presented next.

2.1. The Generalized Extreme Value (GEV) Distribution

To model the extreme values using the GEV distribution a series of N independent observations $x_1, x_2, \dots, x_N$, first blocked into m blocks of size n with n reasonably large and hence N=mn. For weather data, the block size is usually one year. Then from each block the maxima or extreme value $M_i,\ i = 1,2, \dots, m$ is selected and this form a series of m annual maxima data to which the GEV distribution family can be fitted.

Suppose the annual maxima $x_1, x_2, \dots, x_n$ are independent and identically distributed (IID) with distribution$F(x)$. Let$M_n = Max\{x_1, x_2, \dots, x_n\}, n \in N$. If there exist sequences of normalizing constants $\{a_n > 0\}\ and\ b_n \in R$ such that:

$$Prob\left(\frac{M_n - b_n}{a_n} \leq x\right) \to F^n(a_n x + b_n) \to G(x) \qquad (1)$$

As$n \to \infty$, where G is a non degenerate distribution function. If the result of (1) hold, the distribution F is said to be in the domain of attraction of the extreme value distribution G, written as $F \in D(G)$. Then G belongs to family of distributions that can be summarized by the GEV distribution and has the distribution function

$$G(x, \mu, \sigma, \xi) = exp\left[-\left\{1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right\}^{-1/\xi}\right] \qquad (2)$$

Where $\{x: 1 + \xi\ (x - \mu)/\sigma > 0\}\ and\ \mu, \sigma > 0\ and\ \xi$ are location, scale, and shape parameters, respectively?

The extreme value analysis using GEV by block maxima method is often wasteful of data, especially when more data on the extremes are available, leading to large uncertainties on

[†]which is commonly referred to as the peak over threshold (POT) method

return level estimates. Unlike the block maxima method, the POT method provides a more efficient use of data. In the POT method first a threshold is chosen, all the data above the threshold are being considered, and thus more than one event per year could be included in the analysis.

2.2. The generalized Pareto distribution (GPD)

Suppose $x_1, x_2 \ldots$ içs a sequence of IID with a continuous distribution F (.). Suppose that $M_n = Max\{x_1, x_2, \ldots, x_n\}, n \in N$ and x denote an arbitrary term of the sequence and that F (.) satisfies the condition in expression (1). Then, for suitably large u, the distribution function of (x-u) condition on x>u, i,e,$P(x - u/x > u)$, can be approximated by the GPD, which has a distribution function of the form

$$H(x) = 1 - \left\{1 + \xi \frac{x}{\sigma^*}\right\}^{-1/\xi}, \quad x > 0 \qquad (3)$$

Where $\xi x/\sigma^* > 0 \; and \; \sigma^* = \sigma + \xi(u - \mu)$

Note that $\mu, \sigma \; and \; \xi$ are the location, scale and the shape parameters, respectively as defined in expression (2). That is, if G(x) is the approximating distribution of the block maxima, then there is a corresponding approximate distribution for threshold exceedances from within the generalized Pareto family with shape parameter ξ equal to that of the GEV distribution but the scale parameter $\sigma^* = \sigma + \xi(u - \mu)$ for any given threshold u.

The distribution function in expression (3) for $\xi = 0$ is interpreted by taking the limit ξ approaching zero that is:

$$\lim_{\xi \to 0} H(x) = 1 - exp\left(-\frac{x}{\sigma}\right)$$

An exponential distribution with parameter$1/\sigma$. The GPD usually expressed as a two parameter distribution as follow:

$$H(x, \sigma^*, \xi) = \begin{cases} 1 - \left(1 + \dfrac{\xi x}{\sigma^*}\right)^{-1/\xi}, & \xi \neq 0 \\ 1 - exp\left(\dfrac{-x}{\sigma^*}\right), & \xi = 0 \end{cases} \qquad (4)$$

Where $\sigma^* > 0, x \geq 0 \;\; for \; \xi \geq 0 \;\; and \;\; 0 \leq x \leq {-\sigma^*}/{\xi} \;\; for \; \xi < 0$

By differentiating the GPD in expression (4) with respect, the x , the density distribution is given by:

$$h(x, \sigma^*, \xi) = \begin{cases} \dfrac{1}{\sigma^*}\left(1 + \dfrac{\xi x}{\sigma^*}\right)^{-1/\xi - 1}, & \xi \neq 0 \\ \dfrac{1}{\sigma^*} exp\left(\dfrac{-x}{\sigma^*}\right), & \xi = 0 \end{cases}$$

2.3. Parameters Estimation

2.3.1. Determination of the threshold:

The choice of threshold is simply a compromise between bias and variance. Indeed, a threshold that is too low leads to a bad approximation of the limit law, which has the

consequence of increasing the bias. On contrary, a threshold that is too high causes a shortage of extreme values and therefore the variance is increased. Two methods are proposed to choose the threshold:

-the mean excess function (mean residual life plot)

Is an experimental method that is based on the mean of the Pareto distribution: given a random variable Y that follows a GPD distribution with parameters$\sigma^* \; and \; \xi$, its mean is given by the following expression:

$$E(Y) = \begin{cases} \dfrac{\sigma^*}{1-\xi}, & \xi < 1 \\ +\infty, & \xi \geq 1 \end{cases}$$

Now suppose that the GPD distribution is a valid model for observations that exceed a certain thresholdu_0, coming from a sequence $x_1, x_2, \dots, x_{N_u}$ of V.A., then:

$$E[X - u_0 / X > u_0] = \begin{cases} \dfrac{\sigma^*(u_0)}{1-\xi}, & \xi < 1 \\ +\infty, & \xi \geq 1 \end{cases}$$

Where $\sigma^*(u_0) = \sigma + \xi(u_0 - \mu)$

Then, if the Pareto distribution is a good approximation by choosing the thresholdu_0, it will be by choosing any threshold u greater than u0. Therefore, we have also

$$e(u) = E[X - u / X > u] = \frac{\sigma^*}{1-\xi} = \frac{\sigma + \xi(u-\mu)}{1-\xi} = \frac{\sigma^*(u_0) + \xi(u - u_0)}{1-\xi}$$

Where $\xi < 1 \; and \; u > u_0$

it seen that $e(u)$ is a linear function of threshold u, for$u > u_0$. It will expect to find an approximately linear graph in u, from the value of u, which provides a suitable model for the data.

-parameters stability

It must be noted that the estimation of the shape parameter ξdoes not depend on the choice of the threshold u and therefore that its value should remain constant regardless the value chosen of threshold$u$. On the other hand, the estimation of the scale parameter σ^* is influenced by the chosen threshold$u$. indeed, we saw that $u \; and \; \sigma^*$ are linked by the following relation:

$$\sigma^* = \sigma + \xi(u - \mu) = \sigma^*(u_0) + \xi(u - u_0)$$

To remedy this, we can define a reparameterized (modified) scale parameter constant of u:

$$\breve{\sigma}^* = \sigma^* - \xi u$$

With this definition, and since ξ is constant as a function of u, the estimator of $\breve{\sigma}^*$ should also be constant.

If the GPD is a reasonable model for the exceedances of a thresholdu_0, then the estimates of the modified scale and the shape parameters should be approximately constant to all threshold greater tou_0.

2.3.2. Estimation of the parameter of the GPD by maximum likelihood method (ML)

Several methods have been used in the literature to estimate the parameters of the GPD distribution. For example, the method of moments by Christopeit (1994), the L-moments method by Hosking, 1990 and Hosking & Wallis, 1997. For more details about this method see Hosking, 1990); the Bayesian method by Smith and Naylor (1987), Lye et al. (1993), Coles and Tawn (2005); and the maximum likelihood method (Smith and Naylor, 1987). Which is the most popular and has the advantage of allowing the addition to the fitting of co-variables (such as trends, cycles or physical variables) (Katz et al., 2002). The last method was used to estimate the parameters of the GPD distribution as follows.

If $x_1, x_2, \dots, x_{N_u}$ are the N_u exceedances of a threshold u, then the likelihood function associated with $x_1, x_2, \dots, x_{N_u}$is given by:

$$l(\sigma^*, \xi, x_1, x_2, \dots, x_{N_u}) = \prod_{i=1}^{N_u} h(x, \sigma^*, \xi)$$

The log likelihood function is given by:

$$logl(\sigma^*, \xi, x_1, x_2, \dots, x_{N_u}) = \begin{cases} -N_u log\sigma^* - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{N_u} log\left(1 + \xi \frac{x_i}{\sigma^*}\right), & \xi \neq 0 \\ -N_u log\sigma^* - \frac{1}{\sigma^*} \sum_{i=1}^{N_u} x_i & , \xi = 0 \end{cases} \quad (5)$$

By differentiating this expression (5) with respect to the two parameters of interest, we obtain a system of two equations with two unknowns $\sigma^* and\, \xi$. It is by solving these equations that we obtain the ML estimates (using numerical methods, e.g. using Newto Raphson Algorithm).

2.4. Return level estimation for GPD

The focus of extreme weather events analysis usually lies not on estimates of the GPD parameters rather on application of the fitted model to estimate other quantities. For example, to estimate the extreme quantiles of the annual maxima temperature because these values gives an estimate of the level of mean annual temperature expected to exceed once, on average, in a given number of years. Suppose x_T be the "T" year return level ,i,e, it is the value occurring once in every T years.

Assumes that a GPD distribution of parameters $\sigma^* and\, \xi$ is an adequate model for the excesses of a threshold u. then for $x > u$, it result:

$$P(X > x/x > u) = \left(1 + \frac{\xi x}{\sigma^*}\right)^{-1/\xi}$$

$$\frac{P(X > x)}{P(X > u)} = \left(1 + \frac{\xi x}{\sigma^*}\right)^{-1/\xi}$$

$$P(X > x) = P(X > u)\left(1 + \frac{\xi x}{\sigma^*}\right)^{-1/\xi}$$

$$P(X > x) = \gamma_u \left(1 + \frac{\xi x}{\sigma^*}\right)^{-1/\xi}$$

Where γ_u is a parameter to estimate.

The level x_T, which is exceeded once on average in T observations, satisfies the following equation:

$$\frac{1}{T} = \gamma_u \left(1 + \frac{\xi x}{\sigma^*}\right)^{-1/\xi}$$

We find that the formula of x_p is given by:

$$x_T = \begin{cases} u + \dfrac{\sigma^*}{\xi}\left[(T\gamma_u)^{\xi} - 1\right], & \xi \neq 0 \\ u + \sigma log[T\gamma_u], & \xi = 0 \end{cases}$$

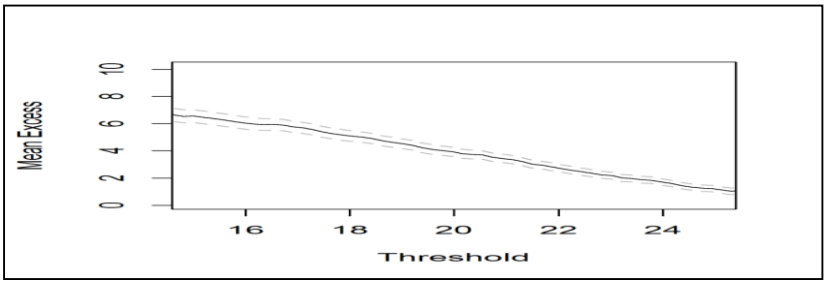
The natural estimator of γ_u is the follow: $\hat{\gamma}_u = \frac{N_u}{N}$, where N_u is the number of exceedances of the threshold u among the N observations. The maximum likelihood estimates of the return level x_p can be obtained using the MLE of σ^* and ξ.

3. APPLICATION

3.1. The choice of threshold

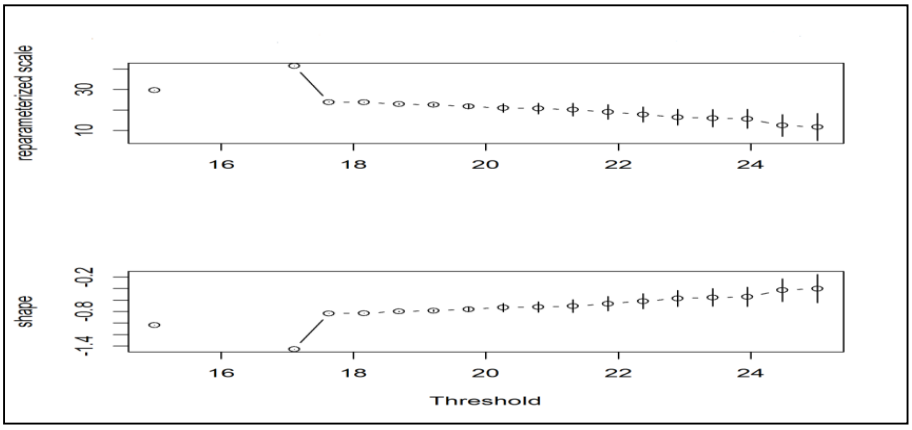
To analyse extreme maximum temperature using the POT method, first a threshold value u_0 is determined and then the GPD is fitted to the temperatures values above u_0. For our data, the threshold value of 18 C° has been chosen using the mean excess plot approach (Figure 1) which is checked by the plots of the ML estimates of the shape and the modified scale parameters against a number of different thresholds (Figure 2).

Figure N° 01: The mean excess plot for the monthly maximum temperature at the jijel weather station



Source: built by myself using R.3.5.2 software

Figure N° 2: Plots of the ML estimates of the shape and the modified scale parameters against a number of different thresholds for the monthly maximum temperature at the jijel weather station

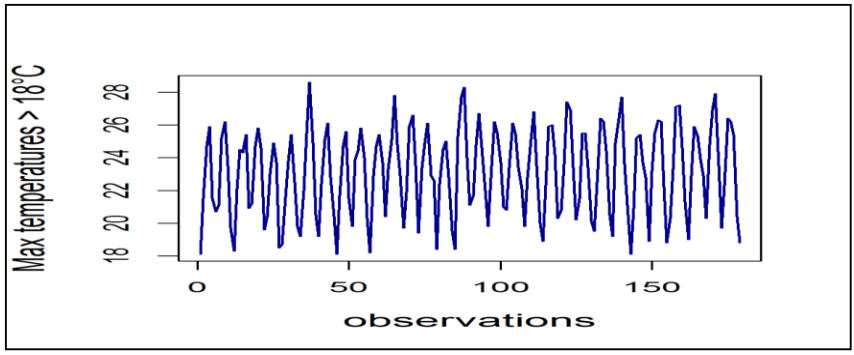


Source: built by myself using R.3.5.2 software

3.2. Descriptive Analysis

The series of the temperatures values above selected threshold $(179\ observations > u_0 = 18\ ^\circ C)$ is given in Figure 3 and its statistical characteristics are given in table 1. From the table it is found that our station is characterized by a mean of 23.14 °C. we can notice that 50% of the data are between 20.90 °C and 25.35 °C also the normality assumption is rejected for the series of the monthly maximum temperatures since the skewness is less than zero, and the kurtosis which is different from 3.

Figure N° 3: Maximum monthly temperatures above 18 °C at weather jijel station



Source: built by myself using R.3.5.2 software

Table N° 01. Statistical properties of the monthly maximum temperature above 18°C at weather jijel station.

N	N_u	Mean	3^{rd}Qu	Max	Skewness	Kurtosis
372	179	23.14	25.35	28.60	-0.17	2.02

Source: built by myself using R.3.5.2 software

3.3. Parameters estimation and model validation

The maximum likelihood method was used (as described in Subsection 2.3.2) to estimate the two parameters of the GP distribution. The Ml estimates of the scale and shape parameters, the associated 95% confidence intervals and the covariance matrix are given in table 2.

Table N° 2. ML estimates, confidence intervals (CI) and covariance matrix of the shape and the scale parameters of GPD model fitted to monthly maximum temperatures.

	Scale	Shape
Estimates	8.83	-0.83
Std.err	0.03	0.00
CI	(8.75, 8.89)	(-0.83,-0.81)
Estimated parameters covariance matrix		

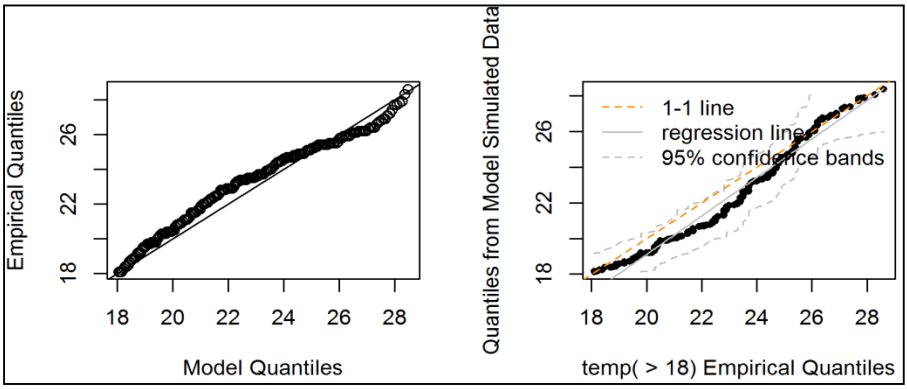
Scale	1.20×10^{-3}	-7.67×10^{-15}
Shape		4×10^{-16}

Source: built by myself using R.3.5.2 software

From Table 2 it is noted that the shape parameter ξ is negative ($\xi = -0.83$) implying that the GP distribution is Pareto type II ; its value is far to zero implying that exponential distribution is excluded, the confidence interval of ξ confirms this conclusion $\xi = 0 \notin CI_{\xi}$.

To validate the chosen model the QQ plot technique was used, it can be seen that for the station the QQ plot is approximately linear; showing that the GPD model type II with threshold 18 ℃ is adequate for the monthly maximum temperatures at the Jijel weather station.

Figure N°04: QQ plot for the Jijel weather station



Source: built by myself using R.3.5.2 software

3.4. Return level estimation for GPD

The estimated return levels, using the ML method for different return periods for the maximum monthly temperatures with 95% profile likelihood confidence intervals (CI) are given in table 3. It can be seen that the return level increase slowly for higher return periods and further the intervals are increasingly wider as the return period is increasing.

Table N° 03. Return levels and 95% IC (in °C) for maximum monthly temperatures using GPD.

Return period	Estimated return level (in °C) CI

2-years	27.22	(20.11 , 34.34)
20-years	28.41	(18.80 , 36.62)
50-years	28.52	(18.03 , 36.77)
100-years	28.56	(17.83 , 36.95)

Source: built by myself using R.3.5.2 software

4. CONCLUSION

In this study the maximum monthly temperatures at the jijel weather station from 1988 to 2018 was modelled using the generalized Pareto distribution (GPD) to control and predict the behaviour of the maxima of temperature. The maximum likelihood method (ML) was used to estimate the parameters and it was found that the Pareto type II (bounded tails) with threshold 18℃ is more appropriate for the jijel weather station. Return levels were estimated for several return time periods; for example, according to this model we must wait about 100 years to record an average monthly temperature of 28.56 ℃.(the temperature value of 28.56℃ is expected to exceed once, on average, in 100 years).

The study of non-stationary model to the maximum monthly temperature or use an alternative method of Bayesian MCMC based on Metropolis Hastings algorithm to estimate the parameter may improve the work.

5. ACKNOWLEDGEMENT

I thank Labeni Fahima and Souyad Hind for their assistance in R software programming and for their technical comments that greatly improved the manuscript.

BIBLIOGRAPHY:

1. **CHRISTOPEIT N**., *« Estimating parameters of an extreme value distribution by the method of moments»*, *J. Stat. Plann. Inference* **41**(2), 1994, PP.173-186.
2. **COLES SG. and TAWN JA.,** *« Statistical methods for multivariate extremes: an speciapplication to structural design»*, Applied Statistics 43, 1994, PP.1–48.
3. **COLES SG . TAWN J.**, « Bayesian modeling of extreme surges on UK east coast », *Philos. Trans. R. Soc. A* **363**, 2005, PP.1387–1406.
4. **COLES SG.,** *«An Introduction to Statistical Modeling of Extreme Values »*, Springer-Verlag: New York, 2001.
5. **DAVISON AC, SMITH RL.,** *« Models for exceedances over high thresholds (withdiscussion) »*, Journal of the Royal Statistical Society, Series B, 1990, PP. 393–442.
6. **EMBRECHTS P. KLUPPELBERG C. and MIKOSCH T.,** *«Modelling Extremal Events for insurance and Finance »*, Springer. New York, 1997.
7. **FISHER RA. And TIPPET LHC**., *« Limiting forms of the frequency distribution of the largest or smallest member of a sample»*, *Proc. Cambridge Philos. Soc.* **24**, 1928, PP. 180–190
8. **FRECHET M**., *«Sur la loi de probabilité de l'ecart maximum »*, *Ann. Soc. Pol. Math.* **6**(3), 1927, PP.93–116.
9. **GILLELAND E. KATZ R W**., « *New software to analyze how extremes change over time*», Transactions of the American Geophysical Union, 92 (2), 2011,PP. 13–14.

10. **GNEDENKO BV**., « *Sur la distribution limite du terme maximum d'une série aléatoire*», *Ann. Math.* **44**, 1943,PP. 423–453.

11. **GUMBEL EJ**., « *Statistics of Extremes* », Columbia University Press, New York, 1958.

12. **HOSKING JRM. And WALLIS JR.,** « *Regional Frequency Analysis. An Approach based on L-Moments*», Cambridge University Press, UK, 1997.

13. **HOSKING JRM.,** *«L-moments: analysis and estimation of distributions using linear combinations of order statistics* » , *J. R. Stat. Soc. Ser. B* **52**, 1990,PP. 105–124.

14. **KATZ RW. PARLANGE MB . and NAVEAU P**., *« Statistics of extremes in hydrology*» , *Adv. Water Resour.* **25**, 2002, PP.1287–1304.

15. **LYE LM, HAPUARACHICHI KP, RYAN S.,** « *Bayes estimation of the extreme value reliability function*», *IEEE Trans. Reliab.* **42**(4), 1993, PP.120-135.

16. **REISS R. and THOMAS M**., «*Statistical Analysis of extreme values* », Birkhauser Verlag, 2001.

17. **SMITH RL. and NAYLOR JC.,** « *A comparison of maximum likelihood and Bayesian estimators for the three parameterWeibull distribution* » , *J. R. Soc., Ser. C* **36**(3), 1987,PP. 358–369.

18. **STEPHENSON AG.,** « *evd: Extreme value distributions. R News*», URL: http://CRAN.R-project.org/doc/Rnews/, 2002, PP.641–644.