Modelling Hydrological Time Series Using Wakeby Distribution

Ilaria Lucrezia Amerise

Abstract—The statistical modelling of precipitation data for a given portion of territory is fundamental for the monitoring of climatic conditions and for Hydrogeological Management Plans (HMP). This modelling is rendered particularly complex by the changes taking place in the frequency and intensity of precipitation, presumably to be attributed to the *global climate change*. This paper applies the Wakeby distribution (with 5 parameters) as a theoretical reference model. The number and the quality of the parameters indicate that this distribution may be the appropriate choice for the interpolations of the hydrological variables and, moreover, the Wakeby is particularly suitable for describing phenomena producing heavy tails. The proposed estimation methods for determining the value of the Wakeby parameters are the same as those used for density functions with heavy tails. The commonly used procedure is the classic method of moments weighed with probabilities (probability weighted moments, PWM) although this has often shown difficulty of convergence, or rather, convergence to a configuration of inappropriate parameters. In this paper, we analyze the problem of the likelihood estimation of a random variable expressed through its quantile function. The method of maximum likelihood, in this case, is more demanding than in the situations of more usual estimation. The reasons for this lie, in the sampling and asymptotic properties of the estimators of maximum likelihood which improve the estimates obtained with indications of their variability and, therefore, their accuracy and reliability. These features are highly appreciated in contexts where poor decisions, attributable to an inefficient or incomplete information base, can cause serious damages.

Keywords—Generalized extreme values (GEV), likelihood estimation, precipitation data, Wakeby distribution.

I. INTRODUCTION

THE traditional procedure that industry researchers use to estimate the order of maximum rainfall size for various recurrence intervals follows the usual method for adapting the most classical models of statistical distributions. The most relevant features are the marked asymmetry, with the values above the median exhibiting a greater weight than those below, and a lengthening towards large values. Positive asymmetry can derive from the presence of a "brake" that becomes activated at a rather low level. Examples of this can be found in an income distribution where reaching a certain level is relatively simple, but greatly exceeding this level is more difficult, or cases of an infectious disease if we consider its progress against the number of days since the outbreak of the epidemic. The source of positive asymmetry in the observations of different phenomena could be due to the simultaneous presence of different situations with respect to relevant variables whose distributions - taken separately - would be symmetrical, but generate asymmetry in their

Ilaria L. Amerise (Dr.) is with the Department of Economics, Statistics and Finance, University of Calabria, Via P. Bucci, 87036, Rende (CS) Italy (e-mail: ilaria.amerise@unical.it).

aggregation. The imbalances concerning the different behavior in the tails are highly relevant, particularly with regard tail thickness and the presence of remote values. An extreme case of positive asymmetry is the "L" curve which is typical of events subject to rarefaction as the number of manifestations is considered. The structure of the paper is as follows. In the next section, we review the properties of the quantile function, which has the merit of representing the behavior of many hydrological variables. Section III analyzes the distribution parameter estimated by Maximum likelihood estimate (MLE). An application to real data is presented in Section IV. The final section summarizes the paper contributions and includes concluding remarks and indicates possible directions for future research.

II. THE QUANTILE FUNCTION

Although they had been known of for many decades, mainly thanks to Italian statistics, [8] introduced several keys to reading the data which, had not yet been framed within a single and consistent project: the estimation of a quantile function (also known as a graduation function) and the quantile density function or sparsity function. According to [8], a quantile production of statistics and analysis of data in a quantile way means placing at the center, not the probabilities or the relative frequencies, but the values of the random variables or the observed modalities.

The quantile function (also known as a tick function) expresses the value of the variable for which the probability p gives the likehood of making an observation that is less or equal to that value and, at the same time, a probability $(1-p)$ of finding a higher value.

$$Q(p) = F^{-1}(p) = \inf\{x \,|\, F(x) \geq p\} \quad \text{per } 0 \leq p \leq 1 \quad (1)$$

where $p = F(.)$ it is the distribution function. Note that the quantile function is symmetrical, if and only if,

$$Q(p) = -Q(1-p) \qquad 0 < p < 1. \tag{2}$$

Quantile estimation is the primary objective of the of hydrological frequencies analysis (See [4]. Among the many models of statistical distributions known in this context, analytical expressions based on quantiles have gained a role of primary importance due to their specific ability to describe the presence of unusual values (*outliers*). The Wakeby quantile function, known by the acronym WAK, has the merit of logical-intuitive representing the behavior of many

hydrological variables over time, [3].

$$X\left(p,\lambda\right)=\lambda_1+\frac{\lambda_2}{\lambda_4}\left(1-q^{\lambda_4}\right)-\frac{\lambda_3}{\lambda_5}\left(1-q^{-\lambda_5}\right);$$

$$0\leq\ p\ \leq\ 1;\quad q=1-p\quad(3)$$

Note that there are other ways in which the Wakeby is given parameters (for example, see [10]). In the version analysed here, the parameters λ_2, λ_3 with $\lambda_2 + \lambda_3 > 0$ are mainly linear, but not exclusively linked to the scale of the variable X_p and each expresses the weight with which the exponentiated component contributes to the formation of value. Finally, λ_4, λ_5 are shape parameters that govern the asymmetry, kurtosis and tails of the density function. As detected by [9], the Wakeby values have a finite and fixed lower threshold equal to $\lambda_1 + (\lambda_2/\lambda_4)$ which is usually constrained to be positive. Note that when $\lambda_4 < 0$ or $\lambda_5 > 0$ the exponentiated components of the WAK are based on the return time $1/(1-p)$ which expresses, a rounded-up value , the number of observations that need to be made to obtain the first value exceeding that of X_p. The lower limit is given by λ_1 in the case where $\lambda_3 > 0$ and $\lambda_5 \geq 0$. The upper limit is $\lambda_1 + (\lambda_2/\lambda_4) + (\lambda_3/\lambda_5)$ if $\lambda_5 < 0$ or if $\lambda_3 = 0$. From another point of view, the WAK model can be obtained from the superposition of three additive components

$$X\left(p,\lambda\right)=c_1+c_2\left(1-p\right)^{\lambda_4}+c_3\left(\frac{1}{1-p}\right)^{\lambda_5}\quad(4)$$

Each component may be traced back to a specific aspect of X_p, which would, thus, result in a mixture of distinct and separate factors. The former acts as a reference level to which the phenomenon would arise in the absence of forces that push it in one direction rather than another. The second is related to the probability that a value above the threshold value is obtained and finally, the third, to the return time. The density function of the WAK, useful for estimation and for graphic representations, can be constructed implicitly starting from the quantile function

$$\frac{1}{\frac{dX(p,\lambda)}{dp}}=h\left[X(p;\lambda)\right]=\frac{1}{\lambda_2}q^{\lambda_4-1}+\lambda_3q^{-(\lambda_5+1)}\quad(5)$$

The parametric regions within which (5) can effectively consider a density function (i.e. positive and with possible asymptotes on the abscissa axis) are the following

$$R_1:\lambda_2+\lambda_3>0,\lambda_2,\lambda_3\geq0$$
$$R_2:\lambda_2>\lambda_3,\lambda_4-1>0,\lambda_5+1<0,\lambda_4<\lambda_5<1,\lambda_4>\lambda_5>1\quad(6)$$
$$R_3:\lambda_2<\lambda_3,\lambda_4-1<0,\lambda_5+1>0,\lambda_5<\lambda_4<1,\lambda_5>\lambda_4>1$$

III. MAXIMUM LIKELIHOOD ESTIMATE (MLE)

The distribution parameter can be conventionally estimated from the available sample data by the method of moments (MOM), maximum likelihood estimator (MLE), probability weighted moments (PWM), or L-moment estimator (LME). Previous studies show that parameter estimates from small samples computed by using the PWM method are less complicated and yet sometimes more accurate than the

MLE method, [6]. Consider a series of observations $\{X_1, X_2, \cdots, X_n\}$ to be represented with a Wakeby distribution. The likelihood function (change of sign) is given by

$$S(\boldsymbol{\lambda})=-\sum Ln\left[h(p_i,\boldsymbol{\lambda})\right]=\sum Ln\left[\lambda_2q_i^{\lambda_4-1}+\lambda_3q_i^{-(\lambda_5+1)}\right];$$

$$q_i=1-p_i\quad(7)$$

where p_i is the solution of the nonlinear equation

$$X_{(i)}=\lambda_1+\frac{\lambda_2}{\lambda_4}\left(1-q_i^{\lambda_4}\right)-\frac{\lambda_3}{\lambda_5}\left(1-q_i^{-\lambda_5}\right)\quad(8)$$

Here $X_{(i)}$ is the observed value occupying the i-th position in the ascending ranking of the $\mathbf{X}$ observations set. It is clear that each evaluation of 7) requires the solution of n nonlinear equations of type 8 compared to p_i which makes the MLE estimation procedure very laborious, but, as will be seen not impractical due to current computing resources. It should be noted that the threshold parameter λ_1 does not explicitly appear in the density function, but has an obvious role in the solution of 8. The $S(\boldsymbol{\lambda})$ criterion could be minimized with respect to $\boldsymbol{\lambda}$ by using the so-called " scoring method ", which is usual for rough iterative likelihood estimates based on the use of the gradient and the Hessian of 7 Of course, we could take out the threshold parameter λ_1 and proceed to the minimization of 7 with respect to the remaining parameters: $(\lambda_2, \lambda_3, \lambda_4, \lambda_5)$, and only after would we reach an estimate for λ_1 solving the simple equation: $\lambda_1 - (\lambda_2 + \lambda_3) = X_{min}$ In the same way we could differentiate the two-step stim procedure: for example by setting the linear parameters $\lambda_1, \lambda_2, \lambda_3$ and using some non-linear methods to search for the optimum only with respect to just the two form parameters λ_4, λ_5. After, if the shape parameters are set, we could use the regression line are for the linear parameters and so on until the convergence of the two phases to a single value of the vector $\boldsymbol{\lambda}$. See, in this regard, [7]. However, we would prefer to set aside these variants and proceed to direct optimization of the likelihood function.

The availability of qualified and efficient computing resources for each user level leads us to adopt an optimization method that avoids the use of first and second derivatives which, due to implicit definitions, would be particularly complex to implement. Among the many possible strategies we prefer the controlled random search method [1], which limits itself to assessing the likelihood function in a 5-dimensional hypercube, progressively replacing better values with existing ones. In general, the extremes between which parameter estimates lie depend on the data being analyzed. The constraints on the parameters have been taken into consideration in each case and the inadmissible configuration is discarded. It is very important to stress that the likelihood function associated with the WAK might have more than one solution (local minima). This is not be surprising if we take into account the high number of parameters present in the model which ensures a considerable degree of flexibility. This, however, is due to the possibility of there being a number of multiple solutions which should not be considered a critical issue; indeed, they could be useful for a better understanding
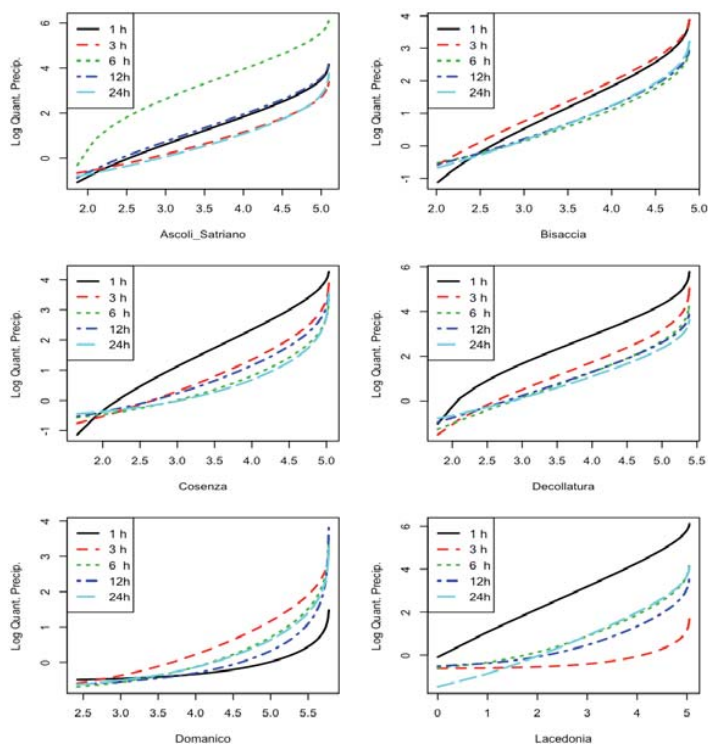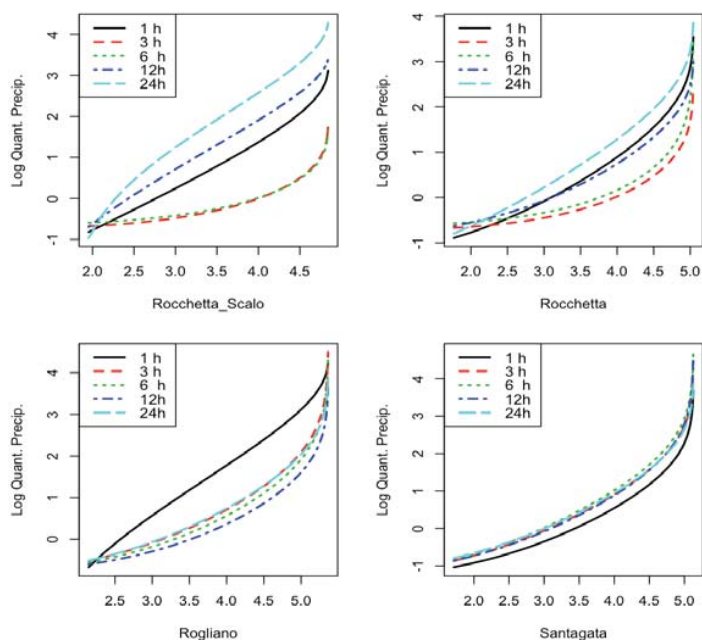
Fig. 1 Wakeby graduation curves

Fig. 2 Wakeby graduation curves

TABLE I
NUMBER OF CASES IN TIME SERIES

Municipality	Start Year	End Year	1 h	3 h	6 h	12 h	24 h
Bisaccia	1959	2008	38	38	39	39	39
Cosenza	1923	2010	61	61	60	61	63
Decollatura	1929	2004	57	55	54	55	54
Domanico	1939	2006	54	54	54	54	54
Lacedonia	1932	2009	41	40	41	40	42
Rocchetta scalo	1947	2008	45	47	46	45	46
Rocchetta	1961	2009	43	43	42	43	44
Rogliano	1961	2009	53	52	51	51	53
Santagata	1930	2008	45	45	44	45	45

of the relationship between the data and the form of the distribution that represents them.

IV. APPLICATIONS

In order to simplify the representation models and to reduce the enormous mass of information, the world-wide organization of the WMO (world metereological organization) suggests the use of thirty year rainfall data. Collections for shorter periods (10-20 years) can be used if the data are treated as cross-sections time series. See [5].

A. Data

Our research refers to the annual time series of hourly averages for a total of 10 weather stations. The length of the time series for the stations differ because they were open for different years and because of interruptions due to specific causes. The range of the values of the time series and the number of observations are shown in Table I.

Only some of the time series are for periods of 30 years, canonically requested to start the interpolation. It should be noted that the time series we have analyzed were subjected to to an efficient system missing values imputation of proposed by [2]. The software, known as the " Amelia " package [5], was entirely developed in the *R* environment, a set of commonly used and free statistical packages which allows simply, rapid application of many multivariate analysis techniques in various disciplines.

B. Results

The figures show the graph for the logarithm of the estimated quantile function of the Wakeby model. For each monitoring station (indicated by the name of the city), the trends for the subdivisions at 1, 3, 6, 12 and 24 hours are presented in the same graph. Attention should be given to the higher curve because it is associated with higher average rainfall levels. The sigmoid appearance which highlights the unimodality of the distribution or the centralization of values around a representative value, is also important. The similar, or even superimposed, shape of the hourly charts indicate little differentiation during the day.

V. CONCLUSION

Statistical frequency distributions are recurrent for the representation of phenomena dominated by a dimensional variable. However, as a consequence of this, the time dimension of the phenomenon is abandoned, ie the information that the values are observations made over time and, therefore, constitute time series, is set aside. The ergodic theorem allows to overcome, at least in part, this perplexity, but it remains a background reserve. We believe that the joint use of models of representation, combined with the analysis of the time series can help to establish the existence (or absence) of a trend of some kind in the water cycle.

REFERENCES

[1] Brachetti P. and Ciccoli, M. and Di Pillo, G. and Lucidi, S. "A new version of the Price's algorithm for global optimization" *Journal of Global Optimization* 10, 165 – 184 (1997).

[2] Dempster, A. P. and Laird, N. M. and Rubin, D. B. "Maximum likelihood from incomplete data via the EM Algorithm" *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1–38 (1977).

[3] Gilchrist, W. G., "Modeling and fitting quantile distributions and regressions", Sheffield Hallam University (2006).

[4] Griffiths, G. A. "A theoretically based Wakeby distribution for annual flood series" *Hydrological Sciences - Journal - des Sciences Hydrologiques* 34, 231–248 (1989).

[5] Honaker, J. and King, G. and Blackwell, M. "Amelia II: A Program for Missing Data" *Journal of Statistical Software* 45, 1–47 (2011).

[6] Jones, R. A. and Scholz, F. W. and Ossiander, M. and Shorack G. R. "Tolerance bounds for log gamma regression models *Technometrics* 27, 109 – 118 (1985).

[7] Lawton, W. H. and Sylvestre, E. A., "Elimination of linear parameters in nonlinear regression" *Technometrics* 13, 461 – 467 (1971).

[8] Parzen, E. "Nonparametric statistical data modelling" *Journal of the American Statistical Association* 74,105 – 131 (1979).

[9] Su, B. and Kundzewicz, Z. W. and Jiang, T. "Simulation of extreme precipitation over the Yangtze River Basin using Wakeby distribution" *Theoretical and Applied Climatology* 96, 209–219 (2009).

[10] Tarsitano, A. "Fitting Wakeby model using maximum likelihood" *Convegno intermedio SIS 2005: Statistica e Ambiente,Messina, 21-23 September, 2005*, 253–256 (2005).