Statistics of extremes through m-component distribution

Carlo COLOSIMO (1) Giuseppe MENDICINO (1) Guo RENDONG (2)

ABSTRACT

By means of a time series analysis of recorded river floods for 19 rivers in Southern Italy, we present a multiple component probabilistic model. Such a model can vary if outliers are present in the considered sample.

The identification of such outliers is carried out by empirical estimation of two typical thresholds, valid for Southern Italian rivers. The combination of these thresholds enables the determination of four distribution classes typified by a number of components (m = 1, 2 and 3) of the probabilistic model under consideration.

KEY WORDS: Extreme floods — Probabilistic model — M-component distribution — Outliers — Southern Italy.

RÉSUMÉ

ÉTUDE STATISTIQUE DES EXTRÊMES PAR LA DISTRIBUTION DES M-COMPOSANTES

Cet article présente un modèle probabiliste à composantes multiples en s'appuyant sur l'analyse des chroniques de crues mesurées sur 19 rivières du sud de l'Italie. Un tel modèle peut être modifié si des points aberrants figurent dans l'échantillon considéré.

L'identification de tels points aberrants est menée au moyen d'une estimation empirique de deux seuils appropriés, valables pour les rivières du sud de l'Italie. La combinaison de ces seuils permet de déterminer quatre classes de distribution caractérisées par le nombre de composantes (m = 1, 2 et 3) du modèle probabiliste.

MOTS CLÉS : Valeurs extrêmes — Modèle probabiliste — Loi à composantes multiples — Points aberrants — Sud de l'Italie.

1. INTRODUCTION

In this century there has been a considerable increase in damage caused by river floods, especially in areas with a large anthropic development. Such an increase depends on both a succession of extraordinary events and structural causes. Natural rivers, depending upon their flow during floods, excavate their own beds in the alluvial plains. If a higher than normal flood occurs, high-flow beds beside of the river are systematically flooded. Human development in these areas is the reason for the continuous growth in damage caused by the floods.

The only effective method of flood-prevention is the realization of passive or active man made protections. However, in the light of the catastrophic events recorded in the past 50 years, such protections have not been of overall effectiveness. In fact, passive intervention typified by strengthening and raising of river banks has reduced the frequency of flooding but is not always capable of solving the problem.

Active structural works, i.e lamination work, are quite effective, but cannot be achieved in areas prone to flooding where human and industrial development is growing.

Thus we do not have absolute security from floods and it is necessary to introduce a risk factor.

Only some of the methods used to estimate flood flow consider the concept of risk. Only statistical methods, based upon either the direct elaboration of the series of the yearly maximum value, X, of instaneous flow, or based

(1) Dipartimento di Difesa del Suolo, Università della Calabria, Contr. da S. Antonello, 87040 Montalto Uffugo (CS), Italia.

(2) Department of Civil Engineering, Shenyang University, Lianhe Road 54, Dadong District, 110041 Shenyang, R.P. China.

upon a longer and more numerous series of the yearly maximum value of the average daily flow (and also upon the series of the values of the flow greater than a predetermined threshold), allow one to estimate the probability of risk corresponding to a particular catastrophic event.

The performance provided by statistical-type procedures varies, however, according to meteorological characteristics which typify a site and, therefore of the corresponding hydrological quantities.

The need to consider a sufficiently uniform methodology for the evaluation of extreme flows is the aim of the present work, which, given the necessary caution deriving from the empirical nature of the study, represents a methodological approach to be verified in other different natural and environmental situations.

2. PROBABILISTIC MODELS

Several probabilistic models provide an estimate, x_T^*, of the theoretical hydrological magnitude, x_T, corresponding to a fixed return period, T.

These models must satisfy the following conditions:

— adequate theoretical basis in order to describe the real process;

— reproduction and explanation of the main statistical features shown by the available samples (values of moments of order greater than one) (VERSACE *et al.*, 1989).

As far as hydrological extremes are concerned, probabilistic models with a theoretical basis follow essentially two different approaches. In the first approach only the maximum value occurring in a fixed time interval (typically one year) is considered, using one value for each year in the series of the yearly maximum values. In the second approach all the values greater than a predetermined threshold are taken into account. The threshold can be modified so that a variable number of yearly events can be considered.

In particular, if a continuous random variable Z is considered, the corresponding cumulative distribution function (CDF), $F_Z(z)$, is the probability that the random variable Z takes a value equal to or smaller than the argument z:

$$F_Z(z) = P\left[Z \le z\right] \tag{1}$$

For the same continuous random variable Z, the probability density function (PDF), $f_Z(z)$, is a function for which the probability that Z lies in the interval (z, z+dz) equals $f_Z(z)\,dz$. This function is represented by the following equation:

$$f_Z(z) = \frac{dF_Z(z)}{dz} \tag{2}$$

Obviously the expected value of the continuous random variable Z is defined as:

$$E\left[Z\right] = \mu = \int_{-\infty}^{+\infty} z\, f_Z(z)\, dz \tag{3}$$

with variance given by:

$$VAR\left[Z\right] = \sigma^2 = \int_{-\infty}^{+\infty} (z-\mu)^2\, f_Z(z)\, dz \tag{4}$$

and skewness equal to:

$$E\left[\left\{(Z - \mu) / \sigma\right\}^3\right] \tag{5}$$

Given n random variables $Z_1, Z_2, Z_3, \ldots\ldots\ldots Z_n$, the maximum is defined as:

$$X = \max Z_i \qquad \text{with } 1 \le i \le n \tag{6}$$

The CDF of X equals by definition:

$$F_X(x) = P\left[X \le x\right] = P\left[Z_1 \le x, Z_2 \le x, \ldots\ldots, Z_n \le x\right] \tag{7}$$

If the random variables Z are independent, then:

$$F_X(x) = P\left[Z_1 \leq x\right] P\left[Z_2 \leq x\right] \ldots\ldots P\left[Z_n \leq x\right] = F_{Z_1}(x)\, F_{Z_2}(x) \ldots\ldots F_{Z_n}(x) \tag{8}$$

If the random variables Z are also all identically distributed with the common CDF $F_Z(z)$, follows:

$$F_X(x) = \left[F_Z(x)\right]^n \tag{9}$$

and if the random variables Z are also continuous with PDF, $f_Z(z)$, we obtain:

$$f_X(x) = n\left[F_Z(x)\right]^{n-1} f_Z(x) \tag{10}$$

In accordance with the preceding definitions, in the first case, one evaluates the possible distribution of the maximum value in a sequence of a large number of independent and identically distributed variables. In the second case, the flood flow is represented as a marked punctual process, and one evaluates the possible distribution of the maximum value of a Poisson-number of random, independent, but not identically distributed variables.

The same considerations should be made regarding the assumption that the original variables, from which is extracted the yearly maximum value, are independent and identically distributed. This hypothesis appears enough restrictive. In fact, it is well known that the successive flows observed in a river section are interdependent, as a result of storage phenomena of the corresponding basin; it is similarly quite rare to have, from one season to another, any variations in the frequency of flows of constant average intensity (identically distributed variables). The interpretation of the sequence of peak flooding by means of a particular stochastic process (called marked punctual process) might indicate a valid alternative to the fundamental problems considered.

If from the chronological diagram of flows $\{Q(t);\ t \geqslant 0\}$ we extract the peak flooding values above a predetermined threshold, q_o, the phenomenon is described by:

— a basic point process, i.e. the sequence of points, τ_i, on the time axis, where the flood above the threshold occurs;

— a sequence of random variables $\{Z_i = Q(\tau_i) - q_o;\ i = 1,2,....\}$ associated with each point i, which represent the excess above q_o and which mark basic punctual process (fig. 1).

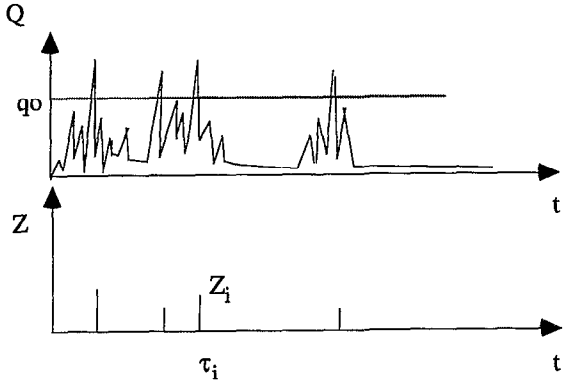


FIG. 1. — Marked punctual process.

If the threshold q_o is high enough, we can assume that the total number of times, K, that the threshold has been exceeded in a fixed time interval follows a Poisson distribution:

$$P\left[K = k\right] = \frac{\Lambda^k}{k!} e^{-\Lambda} \tag{11}$$

where $\Lambda = E[K]$ is the Poisson-process parameter. One assumes the probability that q_o be exceeded at a specific instant, t, is independent of the number of times the threshold has been previously exceeded (CUNNANE, 1979; SALAS *et al.*, 1988; THOM, 1959; TODOROVIC, 1978; TODOROVIC and YEVJEVICH, 1969). As far as floods are concerned, the process follows a Poisson distribution, since catastrophic events, greater than q_o, are far apart from each other in time and so it can be hypothesized their independence.

For this reason, if the threshold q_0 is sufficiently high, if K has a Poisson distribution and the random variables are independent and identically distributed (as well as independent of K), the yearly maximum $X = \max Z_i$ is distributed as follows:

$$F_X(x) = e^{-\Lambda[1 - F_Z(x)]} \qquad x \geq q_0 \tag{12}$$

with a discontinuity of $x = q_0$ equal to $\exp(-\Lambda)$.

Even for $q_0 = 0$ it is possibile to define a yearly number, K, of peak flood events, Z, such that K follows a Poisson distribution and the Z's are independent, identically distributed and independent of K.

Such distribution law has a form which depends upon $F_Z()$. If $F_Z()$ is exponentially distributed and $q_0 = 0$ we find:

$$F_Z(z) = 1 - e^{-z/\theta} \qquad z \geq 0 \tag{13}$$

where $\theta = E[Z]$.

Thus one finds:

$$F_X(x) = e^{-\Lambda e^{-x/\theta}} \qquad x \geq 0 \tag{14}$$

where θ and Λ are respectively $1/\alpha$ and $e^{\varepsilon\alpha}$, typical of GUMBEL'S law (1958):

$$F_X(x) = e^{-e^{-\alpha(x - \varepsilon)}} \qquad \alpha > 0 \tag{15}$$

The time series of the hydrological variable under study, often has several outliers which are considerably different from all the other observed values. Since the number of outliers can be large, we cannot avoid to consider them when we want to estimate the flood flow (BEARD, 1974). This statement is stressed since the corresponding probability of exceeding the extremes in n years can be sensibly different from zero.

The observed distribution law, provides a good estimation of the hydrological variables under study, but it does not give a correct interpretation of these outliers.

In the case of Gumbel's law, the outliers observed in a region in a period of n = 50-60 years, would not be performed since the exceeding probability (in n years) is very close to zero (ROSSI and VERSACE, 1982).

Further improvements can be obtained by means probabilistic models which analyse the series value greater than a predetermined threshold, for seasonal time periods (hypothesizing that on average extreme events occur in the same time of year). Consequently, the distribution function of the annual maximum, X, can be expressed as the product of the distribution functions of maximum, X_1, in the period and of the maximum, X_2, in the other months (TODOROVIC and ROUSSELLE, 1971). Also in this case extreme events may occur in different months from the considered period, as well as floods considered extraordinary in the same period. It can be assumed, therefore, that the two components of random variables $\{Z_i = Q(\tau_i) - q_0;\ i = 1,2,....\}$ cannot be separated completely according to the season: rather, they are mixed opportunely in each of the seasons (ROSSI *et al.*, 1984).

The model TCEV, or also, the two-component extreme value distribution, represents the distribution of the maximum value, in a given time interval, of a random variable distributed according to the mixture of two exponentials:

— the low component (low intensity and high frequency);
— the high component (high intensity and low frequency).

$$F_Z(z) = p\,F_{Z_1}(z) + (1 - p)\,F_{Z_2}(z) \qquad z \geq 0 \tag{16}$$

in which the indices 1 and 2 refer respectively to the low component and to the high one, where p represents the weight of the former component in the mixture.

Analogously, the number of annual excesses, K_1 and K_2, relative to the two components follows a Poisson process, with parameters equal to $\Lambda_1 = E[K_1]$ and $\Lambda_2 = E[K_2]$. On the basis of the « reproductive properties » of the Poisson processes, the comprehensive number of annual exceedings, $K = K_1 + K_2$, also follows a Poisson process with the parameter, $\Lambda = \Lambda_1 + \Lambda_2$. Then, we find:

$$p = \frac{\Lambda_1}{\Lambda_1 + \Lambda_2} \qquad ; \qquad (1 - p) = \frac{\Lambda_2}{\Lambda_1 + \Lambda_2}$$

Definitively, the probability distribution law TCEV of the maximum value X, in a time interval equal to one year, with the function $F_Z()$ exponentially distributed and $q_0 = 0$, results:

$$F_X(x) = e^{\left[-\Lambda_1 e^{-\frac{x}{\theta_1}} - \Lambda_2 e^{-\frac{x}{\theta_2}}\right]} \qquad x \geq 0 \tag{17}$$

having indicated $\theta_1 = E[Z_1]$, $\theta_2 = E[Z_2]$, Λ_1 e Λ_2 the distribution parameters, with:

$$\theta_2 > \theta_1 > 0 \; ; \; \Lambda_1 > 0 \; ; \; \Lambda_2 \geq 0$$

If the number of outliers present in the sample is increased, or if their extraordinary nature is particularly pronounced (meteorological conditions typical of Mediterranean regions) it seems evident how advantageous, on the one hand, it is to consider more components of the random variable within the distribution law and, on the other, not to bind the said law to the only product of two exponential functions. Neither should one neglect those observed value samples typified by a complete absence of outliers, which therefore would not justify the use of a double-component probability distribution law.

The idea of using a more general law of probability distribution (MCEV) type:

$$F_X(x) = e^{\left[\sum_{j=1}^{m} -\Lambda_j e^{-\frac{x}{\theta_j}}\right]} \qquad x \geq 0 \tag{18}$$

which, for m = 1 coincides with Gumbel's Law, for m = 2 represents the TCEV model, and for increasing m values provides a more flexible model for the statistical reproduction of the extraordinary events present in the sample. It would represent a possible alternative to the need to identify a valid distribution law, by means of a uniform model, for samples typified both by the presence and absence of outliers.

3. MCEV MODEL

The multiple-component probability distribution law expressed by equation (18), presents 2m parameters whose calculation can be obtained by the method of maximum likelihood (FIORENTINO and GABRIELE, 1984). Such a method presents a system of equations which are solved by means of an iterative scheme.

In particular, density of probability function is considered:

$$f_X(x) = F_X(x)\Psi_X(x) \tag{19}$$

where $F_X(x)$ is expressed by the equation (18), while $\Psi_X(x)$ is obtained by means of the equation:

$$\Psi_X(x) = \sum_{j=1}^{m}\left[\frac{\Lambda_j}{\theta_j} e^{-\frac{x}{\theta_j}}\right] \tag{20}$$

the natural logarithm of the function of likelihood, L, becomes:

$$L = \sum_{i=1}^{n} \ln f_X(x_i)$$

or :

$$L = \sum_{i=1}^{n} \ln F_X(x_i) + \sum_{i=1}^{n} \ln \Psi_X(x_i) \tag{21}$$

whose partial derivatives with respect to the parameters of distribution, equalized to zero, are:

$$\frac{\partial L}{\partial \Lambda_j} = -\sum_{i=1}^{n} e^{-x_i/\theta_j} + \frac{1}{\theta_j}\sum_{i=1}^{n}\left[\frac{e^{-x_i/\theta_j}}{\Psi_X(x_i)}\right] = 0 \qquad j = 1,2,.....,m \tag{22}$$

$$\frac{\partial L}{\partial \theta_j} = -\frac{\Lambda_j}{\theta_j^2}\left\{\sum_{i=1}^{n} x_i e^{-x_i/\theta_j} + \sum_{i=1}^{n}\left[\frac{e^{-x_i/\theta_j}\left(1-\frac{x_i}{\theta_j}\right)}{\Psi_X(x_i)}\right]\right\} = 0 \qquad j = 1,2,.....,m \tag{23}$$

Multiplying both the members of equations (22) for Λ_j and solving in function of θ_j the equations (23) we obtained by means of an iterative convergent for successive substitutions scheme, the following parameters:

$$\Lambda_j = \Lambda_j \frac{\left[\sum_{i=1}^{n} \frac{e^{-x_i/\theta_j}}{\Psi_X(x_i)}\right]}{\theta_j \sum_{i=1}^{n} e^{-x_i/\theta_j}} \qquad j = 1,2,.....,m \tag{24}$$

$$\theta_j = \frac{\left[\sum_{i=1}^{n} \frac{x_i\, e^{-x_i/\theta_j}}{\Psi_X(x_i)}\right]}{\sum_{i=1}^{n} x_i\, e^{-x_i/\theta_j} + \sum_{i=1}^{n} \frac{e^{-x_i/\theta_j}}{\Psi_X(x_i)}} \qquad j = 1,2,.....,m \tag{25}$$

Through the use of a computer code in C language, based upon the following iterative scheme:

$$\begin{aligned} \theta_j(s+1) &= f(\theta_i(s), \Lambda_i(s)) \\ \Lambda_j(s+1) &= f(\theta_i(s+1), \Lambda_i(s)) \\ j = 1,2,.....,m \; ; \; i &= 1,2,.....,m \end{aligned} \tag{26}$$

the parameter values for m = 1, 2 and 3 are determined.

This scheme has supplied convergent solutions in the totality of examined cases.

The use of equation (18) implies the knowledge of a number of components, m, hypothesized in the distribution law. Such a value, obviously becomes more consistent when the nature of the elements of the considered series is extraordinary.

4. OUTLIERS ESTIMATES

The estimation of outliers present within a sample of observed data, or those data points that depart significantly from the trend of the remaining data, has always represented a controversial problem owing to the need to consider them, or otherwise, in the elaboration of a time series. It therefore implies the need to understand whether such extraordinary values represent anomalous values and should hence be disregarded in probabilistic elaborations of a time series, or whether they could be considered, instead, as representative elements of the behaviour of the series and as such, considered in the data analysis.

A valid response to this question has been provided by BEARD (1974) who, by means of a study carried out on 300 annual peak flow series in the USA, has shown the need to consider all the extraordinary values in data analysis, since their exclusion would produce less accurate results than those in the presence of outliers.

The need to take into account outliers in data analysis is confirmed. In fact, one should consider what happens in Southern Italy where the maximum rainfall and the maximum flow values significantly greater than all other values observed at the same measurement station usually correspond to great downpours which have occurred in the last fifty years.

Obviously such extreme events cannot be considered as isolated and exceptional, but rather, owing to their frequency, should be introduced into the statistical analysis of maximum flows.

The absence of objective criteria for treating outliers, requires empirical evaluations based on both mathematical and hydrological considerations.

In this sense the most used method in the USA and proposed by the Water Resources Council (1981), recommends that if the station skew is greater than 0.4 the following frequency equation can be used to detect high outliers:

$$y_H = y_m + K_n s_y \tag{27}$$

where y_H is the high outlier threshold in log unit, y_m is the mean of log-transformed X values, s_y is the standard deviation and K_n is a coefficient tabled for sample size n. The K_n values are used in one-sided tests that detect outliers at the 10-percent level of significance in normally distributed data (CHOW *et al.*, 1988).

If the logarithms of the values in a sample are greater than y_H, then they are considered high outliers. This procedure, however, provided for Southern Italian rivers, time series without outliers.

Since these results do not agree with the greater part of the hydrological studies carried out on the rivers under examination, a different method to single out the outliers should be searched.

According to KOTTEGODA (1984, 1985), if a sample of n flood observations at a particular station, ranked in order beginning with the lowest, is generated by a model with a specified probabilistic distribution, and assuming such distribution as normal (or when the variable can be transferred to normal), the studentized deviate provides a robust statistic for detecting one or more outliers in a sample. In fact, for every observation of the ranked sample it is possible to determine a parameter, B_j, given by:

$$B_j = \frac{\left| x_{n-j+1} - \overline{x}_{n-j+1} \right|}{s_{n-j+1}} \tag{28}$$

in which:

$$\overline{x}_{n-j+1} = \frac{\sum_{i=1}^{n-j+1} x_i}{n-j+1} \quad ; \quad s_{n-j+1} = \sqrt{\frac{\sum_{i=1}^{n-j+1} (x_i - \overline{x}_{n-j+1})^2}{n-j}}$$

are the mean and standard deviation, respectively, for:

$$j = n-1,.....,1 \quad \text{and} \quad x(1) < x(2) < .... < x(n-j+1)$$

such as shown in fig. 2.

For the rivers of Southern Italy it has been empirically observed that the difference $(B_j - B_{j-1})$ produces useful information on the quantitative characterization of the outliers. In that sense, (j-1)th element of the series, with an associated value of B_{j-1}, can represent the threshold value, past which the outliers in the sample are registered, if the difference $(B_j - B_{j-1}) > 1$ is maximum.

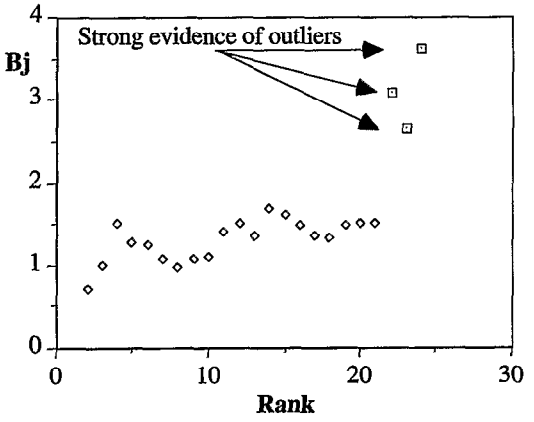


FIG. 2. — Bj's distribution with evidence of outliers.

5. DETERMINATION OF THE M VALUE

Through analysis of the time series of Southern Italian rivers, it is observed that the threshold, $(B_j - B_{j-1}) > 1$, if considered apart, does not allow the determination of the optimum number of components in the probability distribution law. In fact, it is observed that some samples, although not containing outliers, are better explained as functions of double component distribution. For each of these samples however, it is observed that the slope of the regression line between the values assumed by B_j (not classified as outliers) and the corresponding rank, presented high values compared with the average of the slope of all the time series analysed. The introduction of a second threshold based on the value assumed by the slope of the regression line between B_j and the corresponding rank, has enabled the definition of an optimum value to be attributed to m for the rivers of Southern Italy.

With regard to the threshold value relative to the slope of the line of regression, this one emerges from the observation of some samples (Sinni River and Coscile River) respectively characterized by values of the slope of the regression curve equal to 0.041 and 0.044. For such a serie if a value of m = 2 is hypothesized, the iterative scheme represented by the equations (26) shows quite a long convergence time. It would seem, therefore, that the value as-

sumed by the slope of the regression curve, equal to 0.045, is a critical value between low distribution characterized by m = 1, and the high one represented by values of m $\geqslant$ 2.

For a given sample hypothesizing to consider, on the one hand, the maximum difference $(B_j - B_{j-1})$ and on the other, the slope of the regression curve between the B_j values (unclassified as outliers) and the corresponding rank, then four classes can be determined each of which is characterised by a specific m value.

In particular we have:

— *Low distribution*, characterised by the maximum difference $(B_j - B_{j-1}) \leqslant 1$, and the slope of the regression curve less than the threshold value, fixed equal to 0.045. The sample which satisfies such conditions does not present outliers and, therefore, can be explained by an MCEV model characterized by a value of m = 1.

— *Low distribution with outliers*, characterized by the maximum difference, $(B_j - B_{j-1}) > 1$, and the slope of the regression curve less than 0.045. In such conditions, the presence of outliers is ascertained, an MCEV model can be hypothesized characterized by a value of m = 2.

— *High distribution*, characterized by the maximum difference $(B_j - B_{j-1}) \leqslant 1$ and the slope of the regression curve greater than 0.045. In this case the sample which satisfies such conditions, although not presenting outliers, is characterized by a value of m = 2.

— *High distribution with outliers*, characterized by the maximum difference, $(B_j - B_{j-1}) > 1$, and the slope of the regression curve greater than 0.045. For this last class, finally, an MCEV model can be hypothesized with a triple component (m = 3).

6. AVAILABLE DATA

The data used regard flood rates of 19 hydrometric stations, run by the S.I.I (Servizio Idrografico Italiano) in the period 1925-1984 (fig. 3).

The observed samples, with variable dimensions between 19 and 50 years, have been de-dimensionalized with respect to the corresponding average rates, x_m, (table I). The probability distribution law of the extreme value with multiple components used, therefore, gave:

$$F_Y(y) = e^{\left[\sum_{j=1}^{n} -\Lambda_j^* e^{-\frac{y}{\theta_j^*}}\right]} \qquad y \geq 0 \tag{29}$$

where:

$$y = \frac{x}{x_m} \quad ; \quad \Lambda_j^* = \Lambda_j \quad ; \quad \theta_j^* = \frac{\theta_j}{x_m}$$

TABLE I
Dimension of observed series and corresponding average rates

Hydrometric stations	Dimension	xm (m³/s)	Hydrometric stations	Dimension	xm (m³/s)
1. Bradano, P.C.	30	184.15	11. Sinni, V.	27	500.93
2. Tacina, R.	25	81.16	12. Esaro, La M.	19	328.84
3. Bradano, T.P.	19	506.11	13. Crati, C.	31	441.42
4. Alaco, M.	19	13.61	14. Alli, O.	47	16.66
5. Amato, M.	26	79.19	15. Ancinale, R.	50	82.35
6. Corace, G.	38	151.65	16. Lao, P. di B.	24	214.31
7. Basento, P.	28	36.87	17. Agri, Le T.	27	84.80
8. Basento, G.	39	361.00	18. Sinni, P.	34	234.24
9. Basento, M.	24	405.63	19. Coscile, C.	29	78.28
10. Agri, T.	31	186.64			

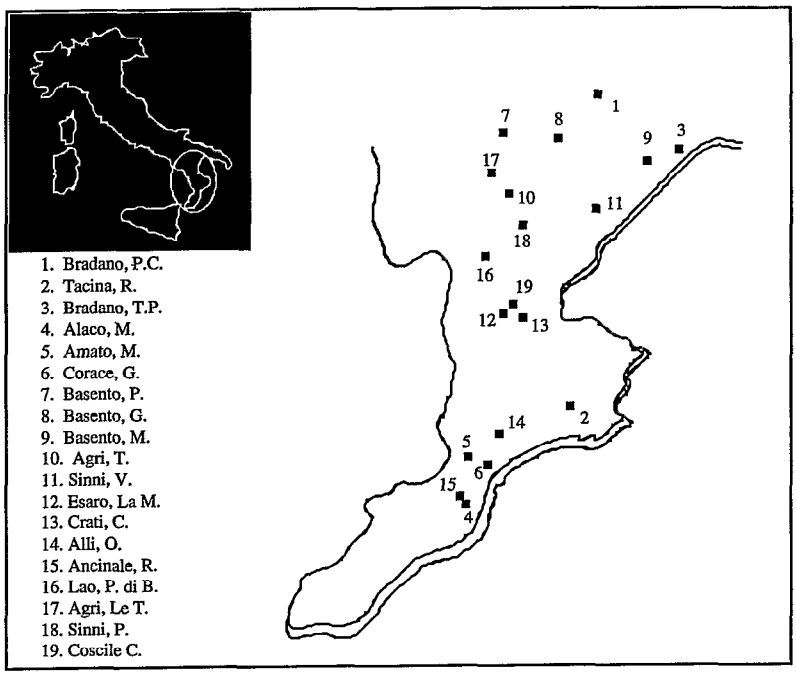


FIG. 3. — Hydrometric stations under examination.

7. RESULTS

For the samples corresponding to the 19 hydrometric stations considered in this paper, we obtained the results in table II.

In particular, in table II the maximum differences ($B_j - B_{j-1}$) are shown, the slope of the regression curve, as well as the classes to which each of the 19 examined series belong.

A representative sample (18. Sinni, P.) which satisfies the required conditions for the class defined as « Low Distribution (m = 1) », is shown in figures 4 and 5. In detail, figure 5 shows a comparison between the accumulated frequencies suggested by Hazen, (i-0.5)/n, of the observed samples and the distribution of probability MCEV for m = 1

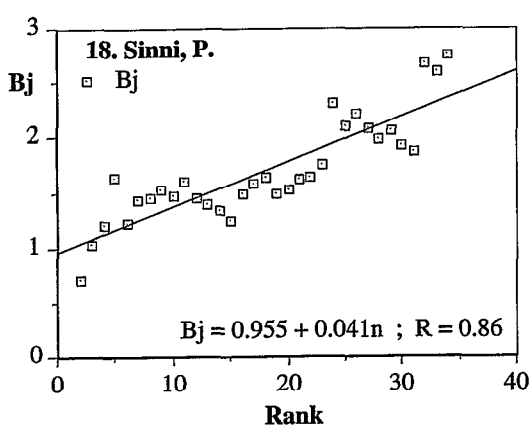
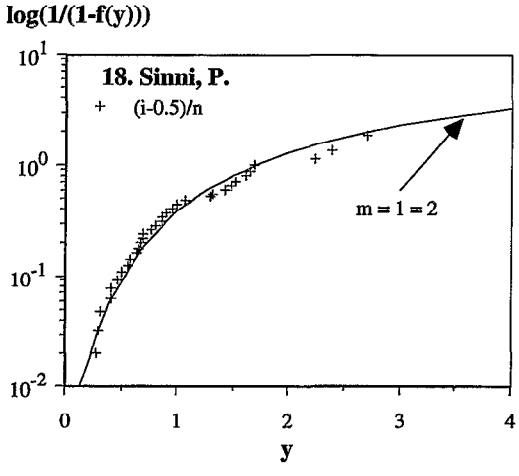


FIG. 4. — Sample which satisfies the required conditions for the class defined as « Low Distribution ».

TABLE II
Classification of samples observed and corresponding to the value of m

Stations	Slope	Max (B_j -B_{j-1})	Class (1-4)	m
17. Agri, Le T.	0.036	0.61	1	1
18. Sinni. P.	0.041	0.80	1	1
19. Coscile. C.	0.044	0.52	1	1
15. Ancinale R.	0.021	1.54	2	2
9 Basento. M.	0.028	1.57	2	2
14. Alli. O.	0.034	1.25	2	2
11. Sinni. V.	0.039	2.06	2	2
6. Corace. G.	0.043	1.15	2	2
8. Basento. G.	0.061	0.49	3	2
13. Crati. C.	0.063	0.70	3	2
7. Basento. P.	0.066	0.95	3	2
10. Agri. T.	0.071	0.77	3	2
16. Lao. P. di B.	0.081	0.46	3	2
3. Bradano. T.P.	0.094	0.77	3	2
4. Alaco. M.	0.102	0.74	3	2
12. Esaro. La M.	0.117	0.84	3	2
5. Amato. M.	0.077	1.17	4	3
1. Bradano P.C.	0.046	1.61	4	3
2. Tacina. R.	0.061	1.01	4	3



m = 1 ---> $\Lambda_1 = 5.271$; $\theta_1 = 0.438$
m = 2 ---> $\Lambda_1 = 3.953$; $\theta_1 = 0.438$
$\Lambda_2 = 1.318$; $\theta_2 = 0.438$

FIG. 5. — MCEV model for values of m equal to 1 (with m = 1 = 2).

A brief consideration must be stressed regarding the quality of the m value proposed. This, if is increased (m = 2), does not cut into the probability distribution, rather only produces an increase in the number of parameters.

Analogous considerations can be made for the other classes. For the representative sample (9. Basento, M.) which satisfies the required conditions for the class defined as « Low distribution with outliers (m = 2) », the results are shown in figures 6 and 7.

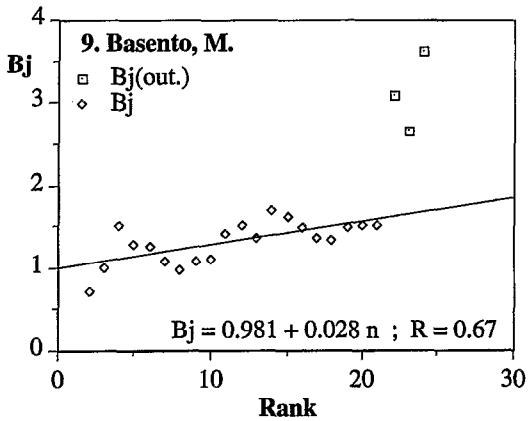


FIG. 6. — Sample which satisfies the required conditions for the class defined as « Low distribution with outliers ».

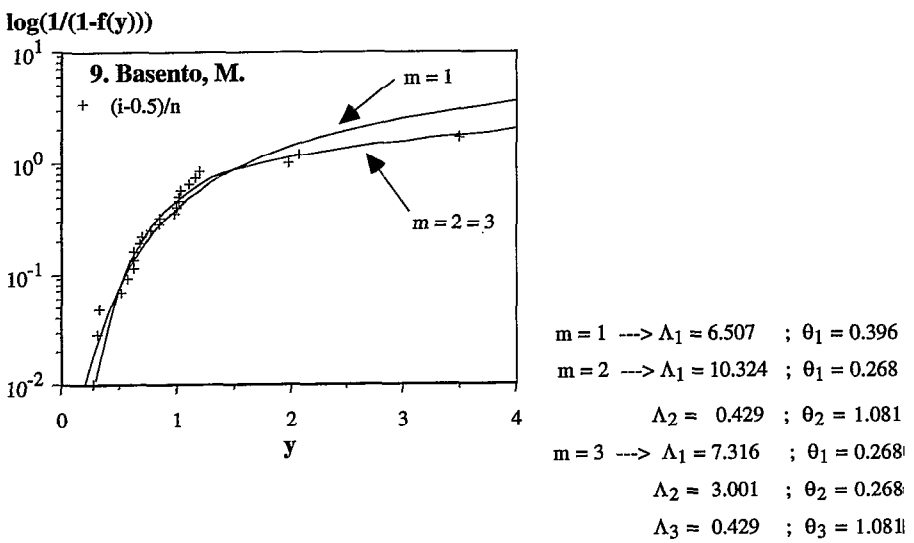


FIG. 7. — MCEV model for values of m equal to 1 and 2 (with m = 2 = 3).

Even in this case, if the m value is increased (m = 3), the probability distribution will not be modified.

For the representative sample (8. Basento, G.) which satisfies the required conditions for the class defined as « High distribution (m = 2) », the results are shown in figures 8 and 9.

For the representative sample (1. Bradano, P.C.) which satisfies the required conditions for the class defined as « High distribution with outliers (m = 3) », the results are shown in figures 10 and 11.

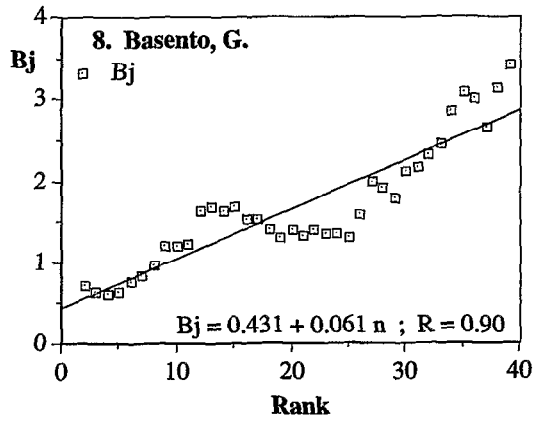
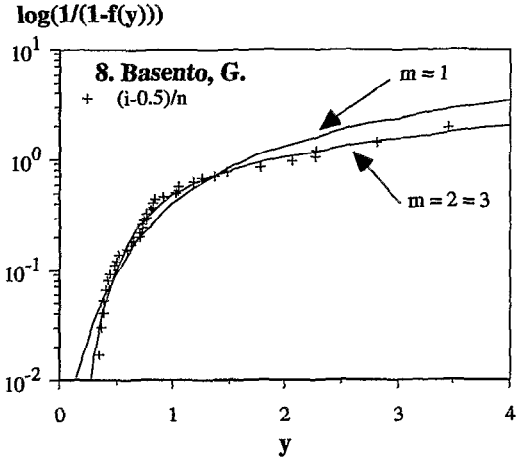


FIG. 8. — Sample which satisfies the conditions required by the class defined as « High distribution ».



m = 1 ---> $\Lambda_1 = 5.306$; $\theta_1 = 0.428$
m = 2 ---> $\Lambda_1 = 11.305$; $\theta_1 = 0.211$
$\Lambda_2 = 0.932$; $\theta_2 = 0.880$
m = 3 ---> $\Lambda_1 = 6.897$; $\theta_1 = 0.211$
$\Lambda_2 = 4.408$; $\theta_2 = 0.211$
$\Lambda_3 = 0.932$; $\theta_3 = 0.880$

FIG. 9. — MCEV model for values of m equal to 1 and 2 (with m = 2 = 3).

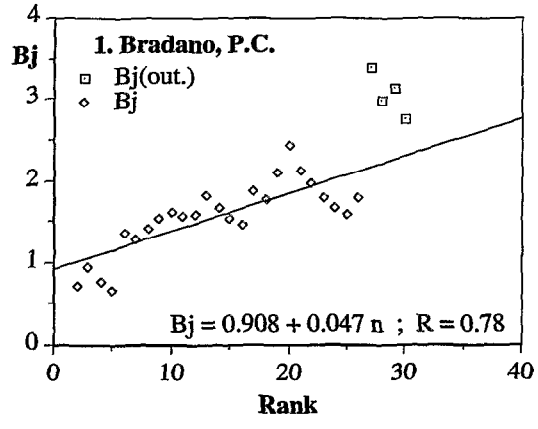
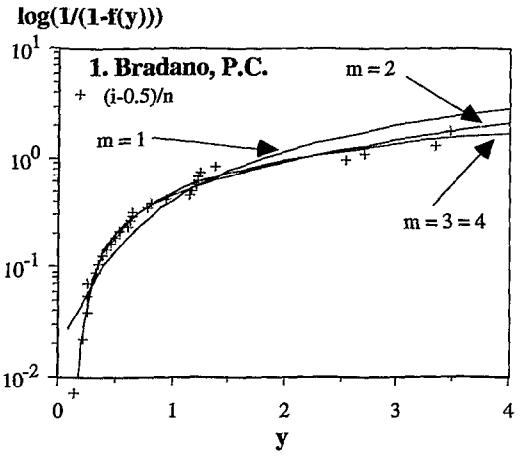


FIG. 10. — Sample which satisfies the required conditions of the class defined as « High distribution with outliers ».



m = 1 ---> $\Lambda_1 = 3.379$; $\theta_1 = 0.528$
m = 2 ---> $\Lambda_1 = 9.789$; $\theta_1 = 0.123$
$\Lambda_2 = 1.537$; $\theta_2 = 0.806$
m = 3 ---> $\Lambda_1 = 21.155$; $\theta_1 = 0.065$
$\Lambda_2 = 2.850$; $\theta_2 = 0.314$
$\Lambda_3 = 0.725$; $\theta_3 = 1.100$
m = 4 ---> $\Lambda_1 = 21.155$; $\theta_1 = 0.065$
$\Lambda_2 = 1.558$; $\theta_2 = 0.314$
$\Lambda_3 = 1.291$; $\theta_3 = 0.314$
$\Lambda_4 = 0.725$; $\theta_4 = 1.100$

FIG. 11. — MCEV model for values of m equal to 1, 2 and 3 (with m = 3 = 4).

8. CONCLUSIONS

By means of a general exponential law, a distribution function of extreme values with multiple components, MCEV, is proposed. Such a model appears particularly flexible to probabilistic interpretation of outliers present in the sample.

Furthermore, it does not reveal potential errors (for excess) in the estimation of values attributable to the number of components, m, of the random variables considered. In fact, in the latter case, high values of m would involve only an increase in the number of parameters, without affecting the quality of the law of distribution.

For the 19 water courses of Southern Italy, in particular, an empirical procedure to classify both the outliers in the samples and the observed samples in homogenous classes characterized by a fixed value of m has been carried out. Finally, for the cases examined, it is demonstrated that the value of m, associated with the generic class is that extremity, going beyond which, the law of distribution does not vary.

The use of such a model, in conclusion, should allow a better prediction of zones at hydrological risk. Such identification need not be exclusively aimed at a more suitable definition of flood protection, rather it should also look at the extreme variability of the rate with regard to risk, with the scope of taking the necessary precautions in the correct hydraulic projection of works.

REFERENCES

BEARD (L. R.), 1974 — *Flood Flow Frequency Techniques*, Technical Report 119, Center for Research in Water Resource, Austin, Texas, University of Texas.

CHOW (V. T.), MAIDMENT (D. R.) and MAYS (L.W.), 1988 — *Applied Hydrology*, McGraw Hill.

CUNNANE (C.), 1979 — A Note on the Poisson Assumption in Partial Duration Models, *Water Resour. Res.*, vol. 15 (2): 489-497.

FIORENTINO (M.) and GABRIELE (S.), 1984 — Distribuzione TCEV: Metodi di Stima dei Parametri e Proprietà Statistiche degli Stimatori, *Geodata*, n° 25, Cosenza.

GUMBEL (E. J.), 1958 — *Statistic of Extremes*, New York, Columbia University Press.

KOTTEGODA (N. T.), 1984 — Investigation of outliers in annual maximum flow series, *J. Hydrology*, 58(1/2): 47-62.

KOTTEGODA (N. T.), 1985 — *Extreme Flood Events and their Effect on Engineering Design*, Advances in Water Engineering, Edited by University of Birmingham UK.

ROSSI (F.), FIORENTINO (M.) and VERSACE (P.), 1984 — Two Component Extreme Value Distribution for Flood Frequency Analysis, *Water Resour. Res.*, vol. 20 (7): 847-856.

ROSSI (F.) and VERSACE (P.), 1982 — « Criteri e Metodi per l'Analisi Statistica delle Piene », in *Valutazione delle Piene*, CNR P.F. Conservazione del Suolo, S.P. Dinamica Fluviale: 63-130.

SALAS (J. D.), DELLEUR (J. W.), YEVJEVICH (J.) and LANE (W. L.), 1988 — *Applied Modeling of Hydrologic Time Series*, Littleton, Colorado, Water Resources Publications.

THOM (H. C. S.), 1959 — A Time Interval Distribution for Excessive Rainfall, *Proc. Soc. Civ. Eng.*, vol. 85, no HY7: 83-91.

TODOROVIC (P.), 1978 — Stochastic Models of Floods, *Water Resour. Res.*, vol. 14 (2): 45-356.

TODOROVIC (P.) and ROUSSELLE (V.), 1971 — Some Problems of Flood Analysis, *Water Resour. Res.*, vol. 7 (5): 1144-1150.

TODOROVIC (P.) and YEVJEVICH (J.), 1969 — Stochastic Process of Precipitation, *Hydrology Paper*, no 35.

U.S. WATER RESOURCES COUNCIL, 1981 — *Guidelines for determining flood flow frequency*, Bulletin 17B, U.S. Geological Survey, Reston, VA 22092.

VERSACE (P.), FERRARI (E.), GABRIELE (S.) and ROSSI (F.), 1989 — *Valutazione delle Piene in Calabria*, CNR-IRPI, Geodata, Cosenza.