

Article

Best Fit and Selection of Theoretical Flood Frequency Distributions Based on Different Runoff Generation Mechanisms

Vito Iacobellis ^{1,*}, Mauro Fiorentino ², Andrea Gioia ¹ and Salvatore Manfreda ²

¹ Dipartimento di Ingegneria delle Acque e di Chimica, Politecnico di Bari, Campus Universitario, Via E. Orabona 4, 70125 Bari, Italy; E-Mail: a.gioia@poliba.it

² Dipartimento di Ingegneria e Fisica dell'Ambiente, Università degli Studi della Basilicata, Via N. Sauro 85, 85100 Potenza, Italy; E-Mails: fiorentino@unibas.it (M.F.); salvatore.manfreda@unibas.it (S.M.)

* Author to whom correspondence should be addressed; E-Mail: v.iacobellis@poliba.it; Tel.: +39-080-5963565; Fax: +39-080-5963414.

Received: 16 March 2010; in revised form: 13 May 2010 / Accepted: 19 May 2010 /

Published: 28 May 2010

Abstract: Theoretically derived distributions allow the detection of dominant runoff generation mechanisms as key signatures of hydrologic similarity. We used two theoretically derived distributions of flood peak annual maxima: the first is the “IF” distribution, which exploits the variable source area concept, coupled with a runoff threshold having scaling properties; the second is the Two Component-IF (TCIF) distribution, which generalizes the IF distribution, and is based on two different threshold mechanisms, associated with ordinary and extraordinary events, respectively. By focusing on the application of both models to two river basins, of sub-humid and semi-arid climate in Southern Italy, we present an *ad hoc* procedure for the estimation of parameters and we discuss the use of appropriate techniques for model selection, in the case of nested distributions.

Keywords: runoff thresholds; frequency distributions; selection criteria; likelihood; model comparison

List of Model Parameters, Units (Parameters without Units are Dimensionless), and Short Description

A (km²): basin area;

τ_A (h): lag-time of basin area A ;

ξ : routing factor;

β : scale parameter of Gamma distribution;

$E[i_{A,\tau}]$ (mm/h): average rainfall intensity referred to the entire basin area A ;

ε : scale parameter of the relationship between average rainfall intensity $E[i_{a,\tau}]$ and source area a ;

q_o (m³/s): base flow;

Λ_p : mean annual number of independent rainfall events;

k : shape parameter of the Weibull distribution of the rainfall intensity;

f_A (mm/h): average hydrologic loss referred to the entire basin area A ;

ε' : scale parameter of the relationship between average hydrologic loss (f_a) and source area a ;

r : ratio of the mean contributing area $E[a]$ to the total basin area A ;

Λ_q : mean annual number of independent flood events;

$f_{A,L}$ (mm/h): lower runoff threshold referred to the entire basin area A ;

$f_{A,H}$ (mm/h): higher runoff threshold referred to the entire basin area A ;

ε_L : scale parameter of the relationship between average hydrologic loss ($f_{a,L}$) and source area a ;

ε_H : scale parameter of the relationship between average hydrologic loss ($f_{a,H}$) and source area a ;

r_L : ratio of the L-type mean contributing area $E[a_L]$ to the total basin area A ;

r_H : ratio of the H-type mean contributing area $E[a_H]$ to the total basin area A ;

Λ_L : mean annual number of independent flood events for L-type;

Λ_H : mean annual number of independent flood events for H-type.

1. Introduction

The identification of dominant processes in flood generation represents the main route for building models able to reproduce real processes and reduce the uncertainty of flood prediction with particular reference to ungauged basins. In this context, the detection of the dynamics responsible for runoff generation and the suitability of a given conceptual scheme may provide interesting insights into basin classification and regionalization.

With this aim, in the recent past, much effort has been spent by hydrologists in order to maximize the exploitation of different kinds of information useful for understanding the hydrological regimes. Basically, in the framework of flood frequency analysis, the uncertainty of prediction is strongly affected by the scarcity of historical data usually due to poor quality or quantity in peak discharge time series. One of the most popular strategies for coping with data scarcity is provided by regional analysis whose purpose is to transfer hydrological information from gauged to ungauged watersheds by identifying hydrologically homogeneous regions and allowing for improved predictions in ungauged basins [1].

In the last few years many studies have been devoted to the analysis of spatial variability in soil properties and land use. They investigate relationships between basin physical properties, model

parameters and hydrological response (assuming that catchments with the same physical characteristics have similar hydrological response) with the aim of finding basin descriptors representative of hydrological signatures (e.g., [2-8]). Others particularly focus on the issue of prediction in ungauged basins [9-13].

Notwithstanding the availability of numerous studies in this field, today no clear guidance is available regarding which model or model structure is appropriate for any particular catchment or management question. Similarly, no clear guidance is available regarding which dominant processes and mechanisms are operating in a given catchment type [14].

A promising opportunity for hydrologists arises from the introduction of physical concepts in the construction of the flood frequency curve by means of derived distributions (e.g., [15-21]). These models, with a simple and physically consistent structure, may provide a valuable compromise between the complexity of real processes and the need for model consistency. In this context, several authors (e.g., [22-24]) have explored the effects due to the coexistence of different runoff processes in flood generation. For instance, Sivapalan *et al.* [22] assumed that floods may be produced by both infiltration excess and saturation excess in the same basin. Another case of coexistence of different runoff processes is given by Allamano *et al.* [24] who proposed a distribution where runoff includes both rainfall and snowmelt contributions.

The “derived distribution” approach provides the opportunity to bridge the gap between purely statistical approaches and physically based (more or less conceptual) simulation models. The first frequently involve the use of distributions that are characterized by many parameters (e.g., [25-29]) and most of them totally lack physical interpretation. On the other hand, advanced knowledge of real processes has driven the construction of several hydrological models used to derive the flood frequency curve based on Monte-Carlo simulations (e.g., [30-32]). These models, in order to achieve reliable predictions, usually require more or less complex procedures for calibration of some key model parameters. In fact, in most cases, direct evaluation of parameters through field observations is not feasible because the scale of measurement is usually much smaller than the effective scale at which the model parameter is applied (e.g., [33,34]).

A novel theoretically derived probability distribution of floods was introduced [20], based on the assumption that two distinct runoff mechanisms are responsible for ordinary and extraordinary flood events. This distribution, called TCIF, is based on the theoretical framework of the IF model (from Iacobellis and Fiorentino [17], where a single runoff mechanism is adopted. The TCIF model generalizes the IF model, for cases which arise when the second component does not exist for return times of technical interest. Comparing models of different complexity, the simpler model is “nested” within the more complex model if it is a special case or restricted version of the other one (e.g., [35,36]). Then, the IF and TCIF models, which are briefly described in Section 2 (with more details in the appendix) are nested distributions. The TCIF distribution was tested in several river basins of Southern Italy characterized by high skewness of the annual maximum flood series (AMFS), providing good performances. Following such results, Gioia *et al.* [20] stated that non-linearity in hydrological processes may be due to the coexistence of different threshold-driven mechanisms of runoff generation. In this paper we investigate basins characterized by less skewed AMFS and dry climatic conditions. We revise the procedure for the estimation of parameters of the IF and TCIF distributions with respect to the previously mentioned applications introducing a different algorithm

for the estimation of the set of parameters which provides the maximum likelihood. More importantly, thanks to the physical meaning associated with the two nested theoretical distributions, we implemented a faster procedure for the estimation of the TCIF parameters which exploits constraints provided by results obtained for the IF parameter values. Section 3 reports the performances of the TCIF and IF distributions in two river basins of Southern Italy, providing interesting insights into the behavior of different runoff thresholds typical of semi-arid and sub-humid climatic conditions.

The issue of model selection for nested distributions is also addressed for the case studies, in Section 4. Although this paper does not aim at exhaustively treating the general problem of model selection, the results obtained are of great interest in the statistical hydrology of extreme events which makes a large use of nested distributions, including, for example, the Generalized Extreme Value (GEV), which generalizes both the Gumbel and Frechet distributions (e.g., [35,37]).

2. Derived Flood Frequency Distributions

2.1. IF Model

The IF model (from Iacobellis and Fiorentino [17]) is based on the concept of partial contributing (or source) area and the peak of direct streamflow Q is considered the product of two random variables strongly correlated, the source area contributing to runoff peak a and the runoff peak per unit of a , u_a , which is

$$u_a = \xi(i_{a,\tau} - f_a), \quad (1)$$

where $i_{a,\tau}$ is the space-time average areal rainfall intensity concerning the contributing area a in the lag-time τ_a and, f_a is the corresponding space-time average hydrologic loss in the area a and in the interval of time τ_a equal to the lag-time of a . The hydrologic loss includes, in general, evaporation of water from the land and vegetative leaf surface, interception of rainfall by vegetation, depression storage on the land surface and infiltration of water into the soil matrix. While considering extreme rainfall-runoff events, here it is mainly referred to infiltration.

The exceedance probability function of the peak of direct streamflow Q , $G_Q'(q)$, is found as the integral of the joint probability density function (PDF) of a and u_a . The IF model assumes that both average rainfall intensity ($E[i_{a,t}]$) and average hydrologic loss (f_a) have important scaling properties:

$$E[i_{a,\tau}] = E[i_{A,\tau}](a/A)^{-\varepsilon} \quad (2)$$

$$f_a = f_A(a/A)^{-\varepsilon} \quad (3)$$

where $E[i_{A,\tau}]$ and f_A are the average rainfall intensity and the average hydrologic loss referred to the entire basin area A . It is useful to remark that introducing the average rainfall intensity ($E[i_{a,t}]$) and the average hydrologic loss (f_a), we apply the $E[]$ operator (expected value) only to rainfall intensity because it is considered a random variable whose entire distribution is exploited in the model. On the other hand, f_a is a quantity that deterministically scales with area and time.

The variable contributing area distribution has parameters α and β which respectively control position and scale, while the following relationship holds:

$$\alpha = rA / \beta \tag{4}$$

where $r = E[a]/A$.

Thus, under the hypothesis that the annual maximum floods arise from a compound Poisson process, Iacobellis and Fiorentino [17] derived the cumulative distribution function (CDF) of the annual maximum flood peak Q_p by means of the relationship:

$$CDF_{Q_p}(q_p) = \exp\{-\Lambda_q [G_Q'(q_p)]\}, \tag{5}$$

where G_Q' is the exceedance probability function of peak flow Q , Λ_q the mean annual number of independent flood events, which is related to the mean annual number of independent rainfall events (Λ_p), the average rainfall intensity ($E[i_{a,t}]$) and the average hydrologic loss (f_A) referred to the entire basin area A :

$$\Lambda_q = \Lambda_p \exp(-f_A^k / E[i_{A,\tau}^k]) \tag{6}$$

The IF probability density function expressed as the derivative of the CDF is:

$$PDF_{Q_p}(q_p) = CDF_{Q_p}(q_p) \left[\Lambda_q \left\{ \int_0^A g(a) \frac{k}{(\xi a) (E[i_{a,\tau}]/\Gamma(1+1/k))^k} \left(\frac{(q_p - q_0)}{(\xi a)} + f_a \right)^{k-1} \exp\left(-\frac{((q_p - q_0)/(\xi a_L) + f_a)^k - (f_a)^k}{(E[i_{a,\tau}]/\Gamma(1+1/k))^k} \right) da \right\} \right] \tag{7}$$

2.2. Two Component IF Model (TCIF)

Gioia *et al.* [20] generalized the IF theoretical probability distribution introducing a two-component derived distribution called “Two Component IF” distribution (TCIF). They identified two different response types, linked to different runoff thresholds, starting from the consideration that different mechanisms may arise, in any basin, with different frequency and magnitude (e.g., [22]). The two different threshold-driven processes are defined as:

- “L-type” (frequent) response, occurring when a lower threshold $f_{a,L}$ is exceeded, and responsible of ordinary floods likely produced by a relatively small portion of the basin a_L :

$$u_{a,L} = \xi (i - f_{a,L}) \quad \text{with} \quad f_{a,L} = f_{A,L} (a_L / A)^{-\varepsilon_L} \tag{8}$$

- “H-type” (rare) response, occurring when a higher threshold $f_{a,H}$ is exceeded, and providing extraordinary floods mostly characterized by larger contributing areas a_H :

$$u_{a,H} = \xi (i - f_{a,H}) \quad \text{with} \quad f_{a,H} = f_{A,H} (a_H / A)^{-\varepsilon_H} \tag{9}$$

The flood-peak contributing areas a_L and a_H are assumed, in analogy with the IF model, as Gamma distributed, with $\beta = 4$, and different mean values.

Therefore, two dimensionless parameters are introduced:

$$r_L = E[a_L]/A \quad \text{and} \quad r_H = E[a_H]/A \quad \text{with} \quad r_H \geq r_L. \tag{10}$$

Assuming that L-type and H-type events are independent and that both rates of occurrence are Poisson distributed, the overall process of exceedances is also a Poisson process and the CDF of the annual maximum floods is

$$CDF_{Q_p}(q_p) = \exp\{-\Lambda_L [G'_{Q,L}(q_p)] - \Lambda_H [G'_{Q,H}(q_p)]\}, \tag{11}$$

where $G'_{Q,L}$ and $G'_{Q,H}$ are the exceedance probability functions of peak flow corresponding respectively to L-type events and H-type events; Λ_L and Λ_H are respectively the mean annual number of independent flood events for L-type and for H-type processes and are related to the runoff thresholds by means of the following relationships:

$$\Lambda_q = \Lambda_L + \Lambda_H = \Lambda_p \exp\left(-\frac{f_{A,L}^k}{E[i_{A,\tau}^k]}\right) \text{ and } \Lambda_H = \Lambda_p \exp\left(-\frac{f_{A,H}^k}{E[i_{A,\tau}^k]}\right), \tag{12}$$

The TCIF cumulative distribution function and its probability density function are:

$$CDF_{Q_p}(q_p) = \exp\left\{-\Lambda_L \left[\int_0^A g(a_L) \exp\left(-\frac{((q_p - q_o)/(\xi a_L) + f_{a,L})^k - (f_{a,L})^k}{(E[i_{a_L,\tau}]/\Gamma(1+1/k))^k}\right) da_L\right]\right\} + \exp\left\{-\Lambda_H \left[\int_0^A g(a_H) \exp\left(-\frac{((q_p - q_o)/(\xi a_H) + f_{a,H})^k - (f_{a,H})^k}{(E[i_{a_H,\tau}]/\Gamma(1+1/k))^k}\right) da_H\right]\right\} \tag{13}$$

$$PDF_{Q_p}(q_p) = CDF_{Q_p}(q_p) \left\{ \Lambda_L \left[\int_0^A g(a_L) \frac{k}{(\xi a_L)(E[i_{a_L,\tau}]/\Gamma(1+1/k))^k} \left(\frac{(q_p - q_o)}{(\xi a_L)} + f_{a,L}\right)^{k-1} \exp\left(-\frac{((q_p - q_o)/(\xi a_L) + f_{a,L})^k - (f_{a,L})^k}{(E[i_{a_L,\tau}]/\Gamma(1+1/k))^k}\right) da_L\right] + \Lambda_H \left[\int_0^A g(a_H) \frac{k}{(\xi a_H)(E[i_{a_H,\tau}]/\Gamma(1+1/k))^k} \left(\frac{(q_p - q_o)}{(\xi a_H)} + f_{a,H}\right)^{k-1} \exp\left(-\frac{((q_p - q_o)/(\xi a_H) + f_{a,H})^k - (f_{a,H})^k}{(E[i_{a_H,\tau}]/\Gamma(1+1/k))^k}\right) da_H\right] \right\} \tag{14}$$

3. Case Studies and Application

In this section, we report results of the application of the IF and TCIF models to two gauged catchments in Southern Italy: the Carapelle river at Carapelle, in Puglia, and the Bradano river at Ponte Colonna, in Basilicata. Puglia and Basilicata are regions in Southern Italy, represented in Figure 1 with the studied basins, their stream network and a 90 m × 90 m digital elevation model (D.E.M.) grid. The main features of the two basins are reported in Table 1, where A is basin area, μ , C_v , C_s and N are, respectively mean, coefficient of variation, coefficient of skewness and sample size of the observed AMFS, I is the Thornthwaite climatic index [38,39], which compares annual precipitation P and annual potential evapotranspiration E_p , $I = (P - E_p)/E_p$. The climatic index distinguishes, in general, between dry ($I < 0$) and humid ($I > 0$) basins. In particular, Carapelle at Carapelle is classified as semi-arid ($-0.4 \leq I < -0.2$) and Bradano at Ponte Colonna as dry-subhumid ($-0.2 \leq I < 0$). They were selected with the aim of finding the most appropriate model structure for each river basin and, consequently, of detecting the presence of different runoff thresholds affecting the processes responsible for runoff generation.

Figure 1. Basins of Southern Italy selected as case studies.

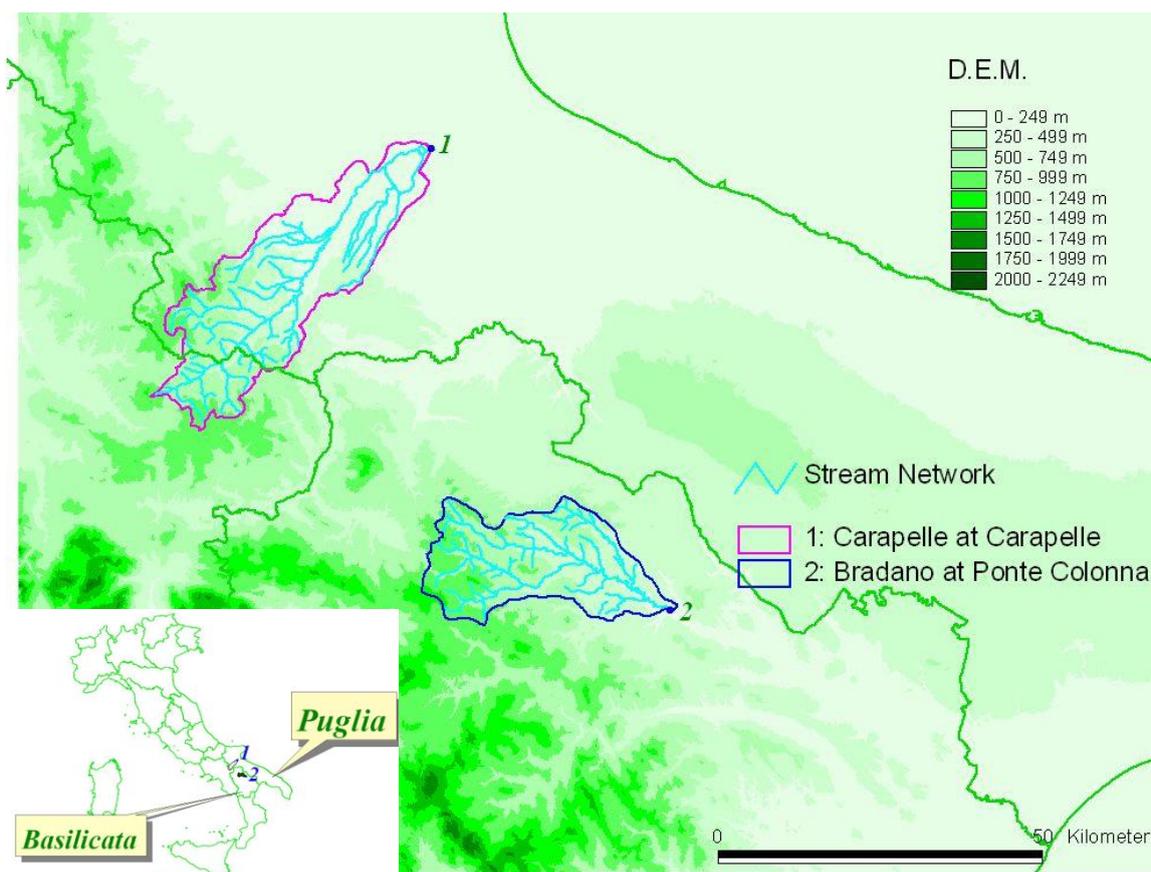


Table 1. River Basin Characteristics.

	n.	A (km ²)	I	(m ³ /s)	Cv	Cs	N
Carapelle at Carapelle	1	715	-0.23	283.7	0.57	1.34	36
Bradano at Ponte Colonna	2	462	-0.08	201.6	0.76	1.21	32

Parameter Estimation and Results

The IF distribution has twelve parameters: baseflow (q_0), four parameters dependent on basin geomorphology (A, τ_A, ξ, β), four rainfall parameters ($E[i_{A,\tau}], \varepsilon, \Lambda_p, k$), and three parameters ($\varepsilon', \Lambda_q, r$), which are strictly related to runoff generation mechanisms. It is worth mentioning that all parameters, with the exception of r and Λ_q , are not calibrated on the available AMFS. We performed for them a *a priori* evaluation by using rainfall statistics and other information. Once all the other parameters are known, only two of them (namely r and Λ_q) are estimated by means of the maximum likelihood function evaluated on AMFS.

The TCIF distribution includes the following fifteen parameters: nine of them are already in the IF model ($q_0, A, \tau_A, \xi, \beta, E[i_{A,\tau}], \varepsilon, \Lambda_p, k$), six more parameters ($\varepsilon_L, \varepsilon_H, \Lambda_H, \Lambda_L, r_L, r_H$) are strictly related to runoff generation mechanisms. Even in this case, four of them ($\Lambda_H, \Lambda_L, r_L, r_H$), are obtained from at-site estimation based on the maximum likelihood function evaluated on AMFS, while all remaining parameters are *a priori* evaluated from information other than AMFS.

We first estimate all parameters of the IF and TCIF distributions which depend on *a priori* information other than AMFS. Among these, all parameters dependent on precipitation which is

analyzed by means of standard regional methods applied to the observed series of annual maxima of rainfall records. Then, the remaining two unknown parameters of the IF model (Λ_q, r) are calibrated using the observed AMFS. Finally, the remaining four unknown parameters of the TCIF model are calibrated by exploiting the IF parameter estimates as initial guess.

Most of the model parameters were estimated in previous studies [17,20,39]. In particular, in Fiorentino and Iacobellis [39], the IF model was applied to several basins in Puglia and Basilicata, including Bradano and Carapelle. Nevertheless we briefly include here the procedures adopted, results are in Table 2. We first describe the evaluation of parameters common to IF and TCIF models. The base flow q_o was estimated as the average monthly flow measured at-site in January and February. There are four parameters dependent on rainfall ($E[i_{A,\tau}], \varepsilon, \Lambda_p, k$) and they were estimated by means of regional frequency analysis of rainfall annual maximum series (AMS) based on the flood index procedure with hierarchical estimation of parameters [39]: k was dependent on the unique regional coefficient of skewness of rainfall AMS; Λ_p was dependent on the regional estimates of the coefficient of variation (different for Basilicata and Puglia). The expected value of the space-time average rainfall intensity $E[i_{A,\tau}]$ was evaluated exploiting the intensity-duration-frequency (IDF) curve of the expected annual maximum rainfall intensity, obtaining the average of the base process from the annual maxima of a Poisson Process and the US Weather Bureau areal reduction factor [20]. The analysis of the regional scaling of $E[i_{A,\tau}]$, provided the regional estimates of the exponent ε (different for Basilicata and Puglia)

Parameters dependent on basin geomorphology are A, τ_A, ξ, β . Basin area A and lag-time τ_A were available in regional studies of basins in Puglia and Basilicata [39], $\beta = 4$ and $\xi = 0.7$ were assigned as described in the appendix (see also [17] for further details).

The loss threshold scaling factors $\varepsilon', \varepsilon_L, \varepsilon_H$ deserve particular attention. For the IF model, ε' was equal to 0.5, assuming that in dry basins the prevalent mechanism is of the storage type [39]. On the other hand, the lower and the higher runoff thresholds of the TCIF model are characterized by Equations (8) and (9), respectively, and particularly by the exponents ε_L and ε_H . In particular, Gioia *et al.* [20] assumed $\varepsilon_L = 0$ and $\varepsilon_H = 0.5$, providing a constant infiltration rate for the lower threshold $f_{a,L}$, and a storage behavior for the higher threshold $f_{a,H}$. With this paper being devoted to the analysis of basins in a dry climate, we assumed for both thresholds a capacitive behavior ($\varepsilon_L = \varepsilon_H = 0.5$) assuming that a dry state characterizes the antecedent soil moisture conditions of both mechanisms. Such an assumption was confirmed by [40] where the analysis was extended to several other basins of Southern Italy.

Table 2. Estimated parameter values of the IF and TCIF models.

Site	q_o (m ³ /s)	$E[i_{A,\tau}]$ (mm/h)	ε	Λ_p	k	τ_A (h)	ξ	β	ε'	Λ_q
Carapelle at Carapelle	7.0	0.20	0.39	44.6	0.8	9.2	0.7	4	0.5	10.5
Bradano at Ponte Colonna	5.0	0.45	0.33	21.0	0.8	4.3	0.7	4	0.5	5.0
Site	f_A (mm/h)	r	ε_L	ε_H	Λ_L	Λ_H	$f_{A,L}$ (mm/h)	$f_{A,H}$ (mm/h)	r_L	r_H
Carapelle at Carapelle	1.01	0.45	0.5	0.5	9.86	0.66	1.01	3.86	0.41	0.99
Bradano at Ponte Colonna	2.02	0.30	0.5	0.5	3.98	1.04	2.01	5.08	0.15	0.99

For the remaining parameters of the IF model, we derived from Equation (6) the relationship

$$f_A = \left[E[i_{A,\tau}^k] \log\left(\frac{A_p}{A_q}\right) \right]^{1/k} \quad (15)$$

providing f_A as a function of A_q , once the *a priori* estimates of k , A_p , $E[i_{A,\tau}]$ are available, and we carried out an at-site evaluation procedure of parameters A_q and r based on minimizing a negative log likelihood function evaluated on AMFS. The procedure was performed by exploring the domain of feasible parameter values on a regular grid. In particular, the grid-dataset values was prepared with r ranging from 0.01 to 1, with step 0.01, and A_q from 0.1 to A_p , with step 0.1. For each test basin, the best parameters dataset was chosen as the one minimizing a negative log likelihood function of the observed sample of annual maximum floods.

Analogously, for the remainder parameters of the TCIF model, we used Equations (12) in order to obtain the expression of $f_{A,L}$ and $f_{A,H}$, using the *a priori* estimates of k , A_p , $E[i_{A,\tau}]$:

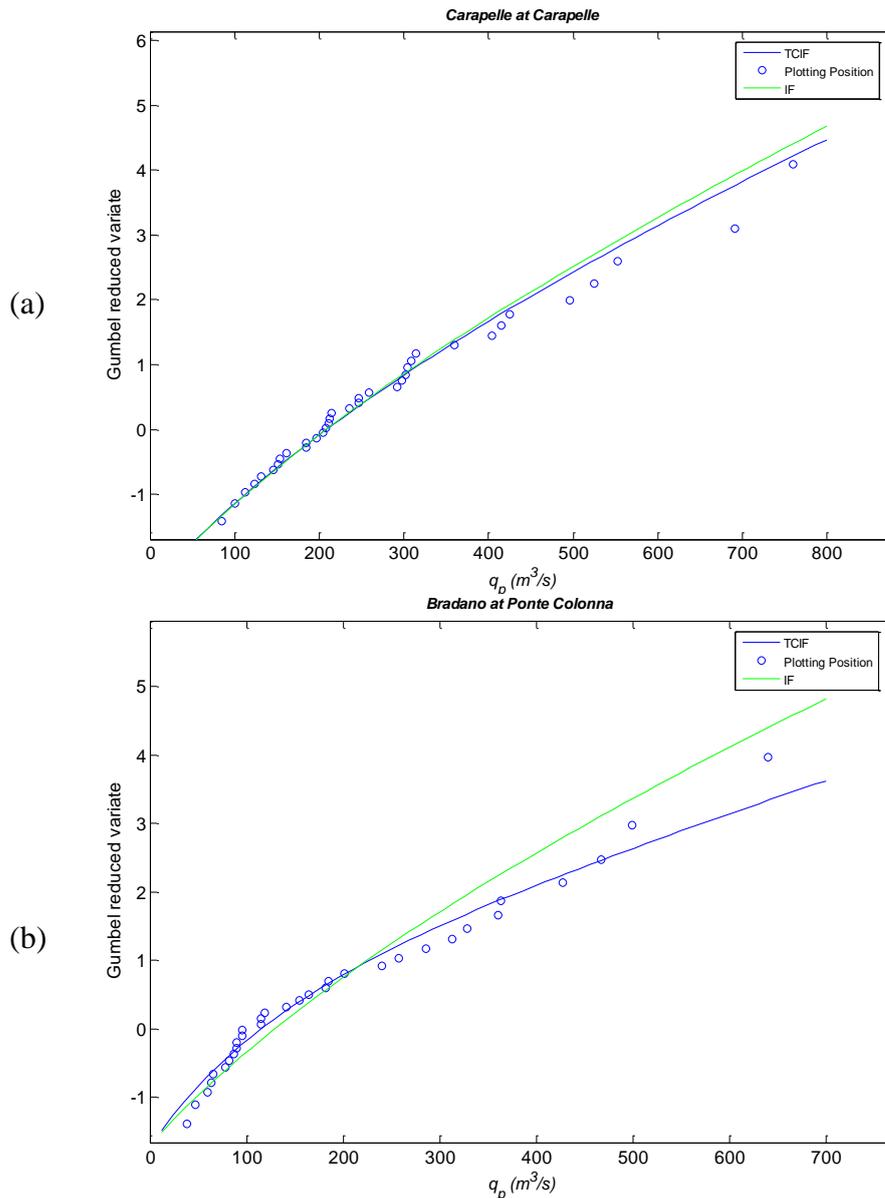
$$f_{A,L} = \left[E[i_{A,\tau}^k] \log\left(\frac{A_p}{A_L + A_H}\right) \right]^{1/k} \quad \text{and} \quad f_{A,H} = \left[E[i_{A,\tau}^k] \log\left(\frac{A_p}{A_H}\right) \right]^{1/k} \quad (16)$$

then, parameters A_L, A_H, r_L, r_H were calibrated adopting a maximum likelihood procedure.

In this case, in order to avoid the cumbersome exploration of the entire parameters domain on a regular grid, we used as initial guess values $r = r_L = r_H$, $A_H = 0$ and $A_L = A_q$. These values correspond to the hypothesis that TCIF distribution collapses into IF distribution. Starting from the initial guess values, the maximum likelihood was found exploring the four-dimensional space of parameters r_L, r_H, A_H, A_L by following the direction of maximum slope of the negative log-likelihood function. In particular, consistently with definitions of the L-type and H-type events we assumed that r_L may only decrease from r to 0, while r_H ranges from r to 1, both with step 0.01; also A_L ranges from A_q to A_p and A_H varies, accordingly (*i.e.*, always respecting the condition $A_L + A_H \leq A_p$) from 0 to A_p . The selected parameter-datasets are reported in Table 2, where it is worth noting that the lower threshold $f_{a,L}$ of the TCIF model corresponds to the single threshold of the IF model. Slight differences are found with respect to the estimates of r and f_A reported in Fiorentino and Iacobellis [39], which used regional estimates of A_q for evaluating f_A , and only r was calibrated by equating an approximate expression of the IF first order moment to the observed mean annual flood.

In Figure 2 we display the TCIF-CDF, the IF-CDF and the Weibull plotting positions of the AMFS of test basins in a Gumbel probability plot. From the visual comparison, one may note the difference between the two distributions which is slight for the semi-arid Carapelle basin while it is more pronounced in the case of the dry-subhumid Bradano basin, the latter being characterized by a more skewed distribution (e.g., [41,42]). In order to objectively assess the “right” model for each test basin, we investigated, as reported in next section, the use of statistical techniques for the “best” model selection.

Figure 2. Comparison between TCIF and IF CDFs and the Weibull plotting positions of the annual maximum flood series: (a) Carapelle at Carapelle and (b) Bradano at Ponte Colonna.



4. Model Selection Procedure

The selection of the best statistical model, for a given sample series, is often based on inference tests depending, for example, on the significance level chosen. In statistical hydrology they are often used in the particular case of nested distributions (e.g., the Gumbel, Frechet and GEV distributions). With the increase of computer capabilities many methods have been proposed and developed for model selection on cross-validation techniques, also in fields other than hydrology (e.g., [43,44]). In this work, five different methods for model selection criteria are used with the aim of finding the most appropriate model structure between the IF and TCIF models.

In particular, we first considered the log-likelihood criterion (LLC), which does not use any penalty term, the Akaike information criterion (AIC) proposed by Akaike [45], who introduced the principle of

maximum entropy for model selection, and the Bayesian information criterion (BIC) proposed by Schwarz [46]. Both AIC and BIC adopt a penalty term accounting for the number of model parameters. The fourth method is the log-likelihood ratio test (LLR). Finally, the *generalization criterion* proposed by Busemeyer and Wang [36]. Mosier [43] was the first to present a clear definition of the cross-validation criterion. Others have shown that the cross-validation criterion is asymptotically equivalent to the AIC [47,48]. Nevertheless, the generalization criterion is based on *a priori* predictions made before observing the data. Thus it objectively assesses the model capability to predict states different from those observed and used for model calibration.

The *LLC* for the *j*th operational model is simply evaluated as:

$$LLC_j = -2 \ln[L_j(\mathcal{G}_j, x)] \quad (17)$$

where $L_j(\mathcal{G}_j, x) = \prod_{i=1}^n g_j(x_i, \mathcal{G}_j)$ is the likelihood function, evaluated at the point $\mathcal{G} = \mathcal{G}_j$, corresponding to the maximum likelihood estimator of the parameter vector \mathcal{G} [49].

The *AIC* for the *j*th operational model may be computed as:

$$AIC_j = -2 \ln[L_j(\mathcal{G}_j, x)] + 2p_j \quad (18)$$

where $L_j(\mathcal{G}_j, x) = \prod_{i=1}^n g_j(x_i, \mathcal{G}_j)$ is the likelihood function, evaluated at the point $\mathcal{G} = \mathcal{G}_j$, corresponding to the maximum likelihood estimator of the parameter vector \mathcal{G} [49], and p_j is the number of estimated parameters of the *j*th operational model. By analyzing Equation (18), one can see that the first term on the right-hand side tends to decrease as more parameters are added to the approximating model, while the second term tends to increase. Note that the penalty term tends to select simpler models under the principle of parsimony. Sugiura [50] derived a second-order variant of AIC, called AICc, valid in the case of a small sample size n , with respect to the number of estimated parameters p ($n/p < 40$):

$$AICc_j = -2 \ln[L_j(\mathcal{G}_j, x)] + 2p_j \left(\frac{n}{n - p_j - 1} \right) \quad (19)$$

BIC for the *j*th operational model is evaluated as follows:

$$BIC_j = -2 \ln[L_j(\mathcal{G}_j, x)] + \ln(n)p_j \quad (20)$$

In practical application one selects the model with the minimum value of the discrepancy measure LLC, AIC or BIC.

Table 3 shows the results of these three selection criteria applied to IF and TCIF. The two criteria (AICc and BIC) accounting for model parsimony suggests rejection of the hypothesis that the AFMS of the two river basins investigated are extracted from the TCIF model which has more parameters, while the negative log-likelihood selection criteria chooses, always, the TCIF model.

Table 3. Application of model selection techniques.

Site	LLC		AICc		BIC		LLR
	IF	TCIF	IF	TCIF	IF	TCIF	(IF, TCIF)
Carapelle at Carapelle	453.12	452.94	457.48	462.23	460.28	467.27	0.18
Bradano at Ponte Colonna	397.29	395.66	401.71	405.14	404.23	409.52	1.63

The log-likelihood ratio test, which is specifically suited for comparisons among nested models, introduces the log-likelihood ratio statistic for two different models i and j :

$$LLR(i, j) = -2 \ln[L_i(\mathcal{G}_i, x) / L_j(\mathcal{G}_j, x)] = LLC_i - LLC_j$$

whose probability distribution can be approximated by a chi-square (χ^2) distribution with $(p_i - p_j)$ degrees of freedom. In this test, if $LLR(i, j)$, where model i is nested within model j , exceeds a cutoff (χ_*^2), which depends on the test significance level, then the null hypothesis that implies no significant model differences (H_0), is rejected. The results in Table 3 show that even the log-likelihood ratio test, which accounts for model parsimony by mean of the chi-square degrees of freedom, always select the IF distribution as being the LLR values well below the cutoff value $\chi_*^2 = 5.99$ obtained for significance level $\alpha = 0.05$ and degrees of freedom = 2.

These results pose a serious question which, for the case of theoretically derived distributions, also has implications on the individuation of the main processes that affect runoff generation. Is it suitable, in this case, to recourse to such selection criteria accounting for model parsimony? An objective answer was found by introducing the generalization criterion which is based on *a priori* predictions made by mean of a split-sample procedure. More precisely, the procedure is structured as follows: (1) For each river basin, the sample of AMFS is divided into two sub-samples statistically independent of sizes N_1 and N_2 (where the total number $N = N_1 + N_2$). (2) During the first calibration stage, best fitting parameter estimates \mathcal{G}_{IF} and \mathcal{G}_{TCIF} are obtained from the sub-sample N_1 , respectively by selecting parameters of the IF and the TCIF models that minimize the discrepancy (D) evaluated as negative log-likelihood function ($D_{IF, N_1} = -2 \ln[L(\mathcal{G}_{IF}, N_1)]$ and $D_{TCIF, N_1} = -2 \ln[L(\mathcal{G}_{TCIF}, N_1)]$) applied to the observed sub-sample N_1 . (3) During the second validation stage the previously estimated parameters, \mathcal{G}_{IF} and \mathcal{G}_{TCIF} , are exploited to compare the two models in terms of their predictive performance calculating the negative log-likelihood function (D_{IF, N_2} and D_{TCIF, N_2}) with respect to the second independent sample N_2 . (4) The difference between the predictive performance of the two models ($\delta_p = D_{TCIF, N_2} - D_{IF, N_2}$) is calculated. (5) Steps 1–4 are repeated 100 times by randomly selecting different sub-samples to produce mean and standard deviation ($\mu(\delta_p)$ and $\sigma(\delta_p)$) of the δ_p factor. If $\mu(\delta_p) > 0$ the IF model is selected and, otherwise, TCIF. The application of this criterion to test basins selected the TCIF model as a better model for both as reported in Table 4 where $\mu(\delta_p)$ and $\sigma(\delta_p)$ are shown.

These results suggest that when cross-validation techniques cannot be applied due to small sample size, the log-likelihood criterion, without any penalty factor accounting for model parsimony, should be preferred when dealing with such nested distributions.

Table 4. Application of generalization criterion for model selection.

Carapelle at Carapelle				Bradano at Ponte Colonna			
N_1	N_2	$\mu(\delta_p)$	$\sigma(\delta_p)$	N_1	N_2	$\mu(\delta_p)$	$\sigma(\delta_p)$
18	18	-0.17	-4.04	16	16	-0.89	-1.61

5. Conclusions

An improvement in flood prediction is expected by selecting the most appropriate model scheme for representing real processes and, consequently, for detecting dominant mechanisms responsible for non-linearity in flood distributions. A comparison between the IF and TCIF models was made using data from two river basins in semi-arid and dry-subhumid climate.

The TCIF model, which generalizes the IF model, introduces two different threshold mechanisms as responsible for ordinary and extraordinary events, in analogy with the theory of the TCEV distribution [26]: The first one is characterized by frequent occurrences and lower average of exceedances (L-type), the second one includes rare events and higher average of exceedances (H-type).

Results of this work show that two different mechanisms of runoff generation may be observed in dry-subhumid and semi-arid climates. In fact, while it is already recognized in the international hydrologic literature that non linear effects in flood frequency distribution may depend on the alternation of infiltration and saturation excess (e.g., [22]), it is less common to observe those different runoff mechanisms in dry climate.

We have shown that a high-frequency behavior may be provided by a storage threshold affecting smaller areas of the basin while the low-frequency component may arise when a higher storage threshold is exceeded in large areas of a basin. Such important results are also compatible with the occurrence of a saturation excess process at the lower component being sometimes modeled as a storage process too (e.g., [22,29]).

We introduced a novel and faster procedure for the estimation of parameters of the TCIF distribution. This is based on the individuation of the maximum-likelihood parameter dataset and is linked to the estimated parameters of the IF distribution, which are used as initial guess while the exploration of the likelihood function is performed respecting the physical constrains $r_L \leq r$, $r_H \geq -r$ and $A_L + A_H \leq A_p$ on the domain of parameter values. Such a procedure is of general interest and may be applied at any climate in any basin where floods are typically rainfall-driven.

Finally, interesting results were obtained in the framework of model selection, in the case of nested distributions IF and TCIF. The comparison was made through the use of selection criteria able to account for the more appropriate model structure. We observed that the selection criteria based on the log-likelihood function, without penalty term, tends to prefer the TCIF model even in the Carapelle basin that does not display a high non-linearity. On the other hand, criteria accounting for a penalty factor related to the number of parameters, such as the AIC, the BIC and the chi-square test, systematically select the IF distribution. This happens also for the dry-subhumid Bradano basin, notwithstanding the TCIF distribution provides a clear better fit to the right tail of the observed distribution (see Figure 2). Then, we referred to the generalized criterion based on a split-sample procedure that objectively tests the *a priori* predictive capability of the model. In both test basins the generalized criterion selected the TCIF distribution, thus providing significant support for its structural

validity and further reinforcing its conceptual representation of the hydrologic processes dominant at basin scale. Such results confirm what was already stated by Busemeyer and Wang [36]: both the Akaike (AIC) information criterion and the Bayesian information criterion (BIC), as well as other methods including a penalty factor, should not be considered appropriate for selecting between nested distributions. They also observed that “it is well known that the chi-square criterion tends to pick the oversimplified model, with small sample sizes that suffer from a lack of statistical power, and it tends to pick the overly complex model in large sample sizes that enjoy extremely high statistical power”. Only the generalized criterion, which is based on cross-validation, performs *a priori* predictions and provides an objective assessment of the model predictive ability. Obviously, the generalized criterion is applicable only to AMFS with minimum length of 30 years, with sub-samples of 15 years used for calibration and validation. Thus, the use of a log-likelihood criterion, without any penalty factor accounting for model parsimony, is recommended when dealing with nested distributions and small sample size. Further investigation on techniques for model selection with diagnostic ability (*i.e.*, able to evaluate model structure validity) is the object of ongoing research and future developments by the authors.

Acknowledgements

The Authors are grateful to two anonymous reviewers for their appropriate and useful comments. The work was realized with support of PRIN—Cubist—CoFin2007 of the MIUR (Italian Ministry of Instruction, University and Research).

References and Notes

1. Sivapalan, M.; Takeuchi, K.; Franks, S.W.; Gupta, V.K.; Karambiri, H.; Lakshmi, V.; Liang, X.; McDonnell, J.J.; Mendiando, E.M.; O’Connell, P.E.; Oki, T.; Pomeroy, J.W.; Schertzer, D.; Uhlenbrook, S.; Zehe, E. IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrol. Sci. J.* **2003**, *48*, 857-880.
2. Abdulla F.A.; Lettenmaier, D.P. Development of regional parameter estimation equations for a macro scale hydrologic model. *J. Hydrol.* **1997**, *197*, 30-57.
3. Seibert, J. Regionalisation of parameters of a conceptual rainfall-runoff model. *Agric. For. Meteorol.* **1999**, *98-99*, 279-293.
4. Fernandez, W.; Vogel R.M.; Sankarasubramanian, A. Regional calibration of a watershed model. *Hydrol. Sci. J.* **2000**, *45*, 689-707.
5. Hundecha, Y.; Bardossy, A. Modeling of the effect of land use changes on the runoff generation of a river basin through parameter regionalization of a watershed model. *J. Hydrol.* **2004**, *292*, 281-295.
6. Merz, B.; Blöschl, G. Regionalisation of watershed model parameters. *J. Hydrol.* **2004**, *287*, 95-123.
7. Wagener, T.; Wheater, H.S. Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty. *J. Hydrol.* **2006**, *320*, 132-154.
8. Heuvelmans, G.; Muys, B.; Feyen, J. Regionalisation of the parameters of a hydrological model: Comparison of linear regression models with artificial neural nets. *J. Hydrol.* **2006**, *319*, 245-265.

9. Post, D.A.; Jakeman, A.J. Predicting the daily streamflow of ungauged watersheds in S.E. Australia by regionalizing the parameters of a lumped conceptual rainfall-runoff model. *Ecol. Model.* **1999**, *123*, 91-104.
10. Wagener, T.; Wheater, H.S.; Gupta, H.V. *Rainfall-runoff Modelling in Gauged and Ungauged Catchments*; Imperial College Press: London, UK; 2004; p. 300.
11. Wagener, T.; Sivapalan, M.; McDonnell, J.J.; Hooper, R.; Lakshmi, V.; Liang, X.; Kumar, P. Predictions in ungauged basins (PUB)—A catalyst for multi-disciplinary hydrology. *Eos. Trans. AGU* **2004**, *85*, 451-452.
12. McIntyre, N.; Lee, H.; Wheater, H.S.; Young, A.; Wagener, T. Ensemble predictions of runoff in ungauged catchments. *Water Resour. Res.* **2005**, *41*, W12434.
13. Young, A.R. Stream flow simulation within UK ungauged watersheds using a daily rainfall-runoff model. *J. Hydrol.* **2006**, *320*, 155-172.
14. McDonnell, J.J.; Woods, R. On the need for catchment classification. *J. Hydrol.* **2004**, *299*, 2-3.
15. Eagleson, P. Dynamics of flood frequency. *Water Resour. Res.* **1972**, *8*, 878-898.
16. Gottschalk, L.; Weingartner, R. Distribution of peak flow derived from a distribution of rainfall volume and runoff coefficient, and a unit hydrograph. *J. Hydrol.* **1998**, *208*, 148-162.
17. Iacobellis, V.; Fiorentino, M. Derived distribution of floods based on the concept of partial area coverage with a climatic appeal. *Water Resour. Res.* **2000**, *36*, 469-482.
18. De Michele, C.; Salvadori, G. On the derived flood frequency distribution: Analytical formulation and the influence of antecedent soil moisture condition. *J. Hydrol.* **2002**, *262*, 245-258.
19. Franchini, M.; Galeati, G.; Lolli, M. Analytical derivation of the flood frequency curve through partial duration series analysis and a probabilistic representation of the runoff coefficient. *J. Hydrol.* **2005**, *303*, 1-15.
20. Gioia, A.; Iacobellis, V.; Manfreda, S.; Fiorentino, M. Runoff thresholds in derived flood frequency distributions. *Hydrol. Earth Syst. Sci.* **2008**, *12*, 1295-1307.
21. Strupczewski, W.G.; Singh, V.P.; Weglarczyk, S. Physics of Environmental Frequency Analysis. In *Integrated Technologies for Environmental Monitoring and Information Production*; Harmancioglu, N.B., Ozkul, S.D., Fistikoglu, O., Geerders, P., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2003; NATO Science Series: IV: Earth and Environmental Sciences, Vol. 23, pp. 195-172.
22. Sivapalan, M.; Wood, E.F.; Beven, K.J. On hydrologic similarity, 3. A dimensionless flood frequency model using a generalized geomorphologic unit hydrograph and partial area runoff generation. *Water Resour. Res.* **1990**, *26*, 43-58.
23. Vander Kwaak, J.E.; Loague, K. Hydrologic-response simulations for the R-5 catchment with a comprehensive physics-based model. *Water Resour. Res.* **2001**, *37*, 999-1013.
24. Allamano, P.; Claps, P.; Laio, F. An analytical model of the effects of catchment elevation on the flood frequency distribution. *Water Resour. Res.* **2009**, *45*, W01402.
25. Waylen, P.; Woo, M. Prediction of annual floods generated by mixed processes. *Water Resour. Res.* **1982**, *18*, 1283-1286.
26. Rossi, F.; Fiorentino, M.; Versace, P. Two-component extreme value distribution for flood frequency analysis. *Water Resour. Res.* **1984**, *20*, 847-856.

27. Buishand, T.; Demaré G. Estimation of the annual maximum distribution from samples of maxima in separate seasons. *Stochastic Hydrol. Hydraul.* **1990**, *4*, 89-103.
28. Alila, Y.; Mtiraoui, A. Implications of heterogeneous flood-frequency distributions on traditional stream-discharge prediction techniques. *Hydrol. Proc.* **2002**, *16*, 1065-1084.
29. Sivapalan, M.; Blöschl, G.; Merz R.; Gutknecht, D. Linking flood frequency to long-term water balance: Incorporating effects of seasonality. *Water Resour. Res.* **2005**, *41*, W06012.
30. Rahman, A.; Weinmann, P.; Hoang, T.; Laurenson, E. Monte Carlo simulation of flood frequency curves from rainfall. *J. Hydrol.* **2002**, *256*, 196-210.
31. Blazkova, S.; Beven, K. Flood frequency estimation by continuous simulation of subcatchment rainfalls and discharges with the aim of improving dam safety assessment in a large basin in the Czech Republic. *J. Hydrol.* **2004**, *292*, 153-172.
32. Fiorentino, M.; Manfreda, S.; Iacobellis, V. Peak runoff contributing area as hydrological signature of the probability distribution of floods. *Adv. Water Resour.* **2007**, *30*, 2123-2134.
33. Beven, K.J. *Rainfall-Runoff Modeling—The Primer*; Wiley: Chichester, UK, 2001.
34. Wagener, T.; Gupta, H.V. Model identification for hydrological forecasting under uncertainty. *Stoch. Environ. Res. Risk Assess.* **2005**, *19*, 378-387.
35. Laio, F.; Di Baldassarre, G.; Montanari, A. Model selection techniques for the frequency analysis of hydrological extremes. *Water Resour. Res.* **2009**, *45*, W07416.
36. Busemeyer, J.R.; Wang, Y.M. Model comparisons and model selections based on generalization criterion methodology. *J. Math. Psychol.* **2000**, *44*, 171-189.
37. Kottegoda, N.T.; Rosso, R. *Statistics, Probability and Reliability for Civil and Environmental Engineers*; McGraw-Hill: New York, NY, USA, 1997.
38. Thornthwaite, C.W. An approach toward a rational classification of climate. *Am. Geograph. Rev.* **1948**, *38*, 55-94.
39. Fiorentino, M.; Iacobellis, V. New insights about the climatic and geologic control on the probability distribution of floods. *Water Resour. Res.* **2001**, *37*, 721-730.
40. Fiorentino, M.; Gioia, A.; Iacobellis, V.; Manfreda S. Regional analysis of runoff thresholds behaviour in Southern Italy based on theoretically derived distributions. *Adv. Geosci.* 2010, (in press).
41. Matalas, N.C.; Slack, J.R.; Wallis, J.R. Regional skew in search of a parent. *Water Resour. Res.* **1975**, *11*, 815-826.
42. Cunnane, C. Review of statistical models for flood frequency estimation. In *Proceedings of the International Symposium on Flood Frequency and Risk Analysis*, Louisiana State University, Baton Rouge, LA, USA, 14–17 May 1986.
43. Mosier, C.I. Problems and designs of cross-validation. *Educ. Psychol. Meas.* **1951**, *11*, 5-11.
44. Camstra, A.; Boomsma, A. Cross-validation in regression and covariance structure analysis: An overview. *Sociol. Method. Res.* **1992**, *21*, 89-115.
45. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*, Tsahkadsor, Armenia, 2–8 September 1971; Petrov, B.N., Csaki, F., Eds.; Akademiai Kiado: Budapest, Hungary, 1973; pp. 267-281.
46. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461-464.

47. Stone, M. On asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Stat. Soc. Ser. B* **1977**, *39*, 44-47.
48. Browne, M.; Cudeck W. Single Sample cross-validation indices for covariance structures. *Multivariate Behav. Res.* **1989**, *24*, 445-455.
49. Linhart, H.; Zucchini, W. *Model Selection*; John Wiley: Hoboken, NJ, USA, 1986.
50. Sugiura, N. Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Stat. Theory Methods* **1978**, *A7*, 13-26.

Appendix

Main features of the IF model are summarized in the following points:

- both random variables a and u_a are controlled by: (i) rainfall intensity, duration and areal extension; (ii) runoff concentration; (iii) hydrological losses.
- The runoff peak per unit area, u_a , is linearly dependent on the areal net rainfall intensity in a time interval equal to τ_a with a constant routing factor ξ . Then, the probability distribution of u_a , can be derived from the probability distribution of rainfall intensity $i_{a,t}$ conditional on a duration equal to τ_a , lag-time of a .
- The areal rainfall intensity $i_{a,t}$ is assumed Weibull distributed with two parameters $\theta_{a,\tau}$ and k . The mean areal rainfall intensity is:

$$E[i_{a,\tau}^k] = \theta_{a,\tau} = (E[i_{a,\tau}]/\Gamma(1+1/k))^k \quad (21)$$

- The routing factor ξ is a key model parameter which in reality appeared very stable. In fact, ξ , it was found to vary in a narrow range (0.6, 0.8) with an average value close to 0.7 which has been used in all the applications of the IF and TCIF models made since they were introduced.
- The lag-time τ_a scales with a according to a power law with exponent 0.5.
- The variable contributing area a follows a mixed distribution with a continuous part which is a two parameter gamma distribution, valid for $0 < a < A$ and a discrete probability P_A

$$g(a) = \frac{1}{\alpha\Gamma(\beta)} \left(\frac{a}{\alpha}\right)^{\beta-1} \exp\left(-\frac{a}{\alpha}\right) + \delta(a-A)P_A \quad (22)$$

$$P_A = \text{prob}[a = A] = \gamma(A/\alpha, \beta) = \int_A^\infty \frac{1}{\alpha\Gamma(\beta)} \left(\frac{a}{\alpha}\right)^{\beta-1} \exp\left(-\frac{a}{\alpha}\right) da \quad (23)$$

The gamma function arises as the distribution of the sum of β stochastic (independent) variables exponentially distributed with equal mean value α .

Thus, being any flood peak due to the superposition of flows coming from sub-basins whose expected number is equal to the number N_ω of sub-basins of Horton order immediately smaller than that of the whole basin, we identified β to $E[N_\omega]$. N_ω tends to be invariant at any scale and assumes values ranging between 3 and 5 [50] with expected value close to 4 [51].

- The annual maximum floods arise from a compound Poisson process and the following relationships hold for the flood peak q_p , the peak of direct streamflow Q , and the exceedance probability function of the peak of direct streamflow $G_Q'(q)$:

$$q_p = Q + q_o, \quad (24)$$

$$G_Q'(q_p) = \int_0^A g(a) \exp\left(-\frac{\left(\frac{(q_p - q_o)}{(\xi a) + f_a}\right)^k - (f_a)^k}{E[i_{a,\tau}^k]}\right) da, \quad (25)$$

with base flow q_o

© 2010 by the authors; licensee MDPI, Basel, Switzerland. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).