

# A new plotting position concept to evaluate peak flood discharges based on short samples

M. M. Portela<sup>1</sup> & J. M. Delgado<sup>2</sup>

<sup>1</sup>*Instituto Superior Técnico (IST), Portugal*

<sup>2</sup>*GeoForschungsZentrum Potsdam (GFZ), Germany*

## Abstract

A model to improve the estimates of the peak flood discharges given by the statistical analysis applied to small samples of annual maximum instantaneous discharges,  $Q_{ami}$ , is presented. The model combines a new probability plotting position technique – by means of a sorting operator, SO – with the Gumbel law with parameters evaluated by the least square method. It should be stressed that other statistical laws compatible with the previous parameter estimation method should be possible. For a given watershed, the SO modifies the rank of each element of the sample of  $Q_{ami}$  by taking into account the magnitude of the floods in a nearby watershed where more records are available. The two watersheds must belong to the same homogenous hydrologic region and must be geographically close in order to be valid to consider that the extreme flood events in both watersheds are likely due to the same extreme rainfall events. The results achieved showed that the model allows for better estimates, especially when applied to short samples, as those more often provided by the Portuguese network of stream gages.

*Keywords: flood analysis, probability plotting position, sorting operator, least square method, Gumbel law.*

## 1 Plotting position concept

The adjustment of a statistical distribution function to a sample of a random variable requires the computation of the empirical non-exceedance probability of each element  $x_k$  of the sample,  $F(x_k)$ , by means of a plotting position formula. For that purpose the values of the sample are ranked, more often from the smallest to the largest and the following formula is applied



$$F(x_k) = \frac{i - \theta}{n + 1 - 2\theta} \quad (1)$$

where  $i$  is the ordered rank of value  $x_k$ ,  $n$  is the sample size, and  $\theta$  is a constant between 0 and 1, depending upon the plotting position formula and on the theoretical distribution (Cunnane [2]). If the sample does not have repeated values – as generally happens with the samples of hydrologic variables –  $i$  is a continuous integer comprehended between 1 to  $n$  ( $1 \leq i \leq n$ ) and it represents the number of elements of the sample equal or smaller than  $x_k$ . In the Weibull, Gringorten and Hazen  $\theta$  is equal to 0.00, 0.44 e 0.50, respectively. In Figure 1 the empirical non-exceedance probabilities given by the three previous formulas applied to two samples composed by 50 (left diagram) and 20 (right diagram) different elements are compared.

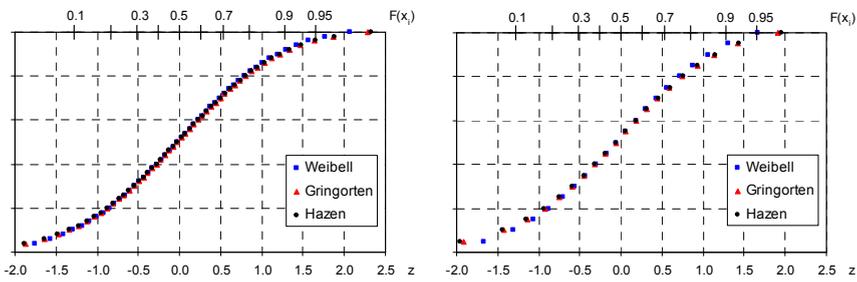


Figure 1: Empirical non-exceedance probabilities given by the Weibull, the Gringorten and the Hazen formulas applied to two samples, one with 50 elements (left side diagram) and the other with 20 elements (right side diagram).

Both diagrams of Figure 1 have two horizontal axes, a principal axis with a linear scale of the standard normal variable,  $z$ , and a secondary axis with a non-linear scale of the non-exceedance probabilities,  $F(x)$ , which correspond to the values of  $z$  according to the standard normal distribution. The figure clearly shows that except for the extreme probabilities the results achieved by the three plotting positions formulas are almost the same.

Aiming at improving the estimates of the peak flood discharges specially when based on small samples of annual maximum instantaneous discharges, Qami (Sample built upon one value per hydrologic year, the maximum instantaneous discharge in that year. In Portugal, the hydrologic year starts October 1<sup>st</sup>), a statistical approach was developed based on a new plotting position concept combined with the Gumbel law with parameters estimates by the least square method, LSM. The new concept allows discontinuous ordering of the values of the sample by means of a sorting operator, SO, that accounts for the relative magnitude of the flood discharges based on a sample of Qami in a nearby watershed where a much larger number of discharge records is available (Delgado [4] and Portela and Delgado [12]).

The SO operator is based on the assumption that the Qami in two neighboring basins are consequence of the same extreme meteorological events. In other words, the floods generated in each of those basins should have similar relative magnitude: the highest floods in both basins should occur in the same period or in vicinal periods, the second highest flood idem, and so on, provided that the basins are geographically close enough and are located in the same homogeneous region regarding the flood generation process (Coursey [1], Dalrymple [3], Dias [5], Gonçalves [6], Lettenmaier *et al.* [10], Loureiro [11] and Portela and Dias [13]).

## 2 Gumbel law and parameter estimation

There is not a single universally accepted model for flood frequency analysis. On the contrary, a whole group of models like lognormal, Pearson III, log-Pearson III, Wakeby or general extreme value distributions has been suggested. However, the type-1 extreme value distribution or Gumbel law (Gumbel [8]), is often applied to the statistical analysis of extreme values such as floods or heavy precipitation (Henriques [9], Raynal and Salas [14] and Reiss and Thomas [15]).

The Gumbel probability density function,  $f(x)$ , and probability distribution function,  $F(x)$ , are represented in Figure 2 where  $\alpha$  ( $\alpha > 0$ ) and  $u$  are the scale and the location (mode) parameters, respectively.

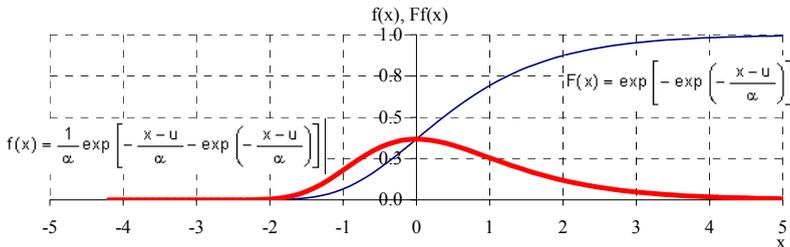


Figure 2: Gumbel probability density function,  $f(x)$ , and probability distribution function,  $F(x)$ : general equations and curves for  $\alpha = 1$  and  $u = 0$ .

Normally, the inverse of the probability distribution function is of more interest, as it allows obtaining the quantile of the cumulative probability  $F(x)$ , i.e., the value,  $x$ , of the random variable,  $X$ , with a given non-exceedance probability:

$$x = u - \alpha \ln[-\ln(F(x))] \quad (2)$$

The first two order moments of this function, the mean,  $\mu$ , and the variance,  $\sigma^2$ , are given by:

$$\mu = u - \varepsilon \alpha \quad \sigma^2 = \pi^2 \alpha^2 / 6 \quad (3) \text{ and } (4)$$

where  $\sigma$  is the standard deviation and  $\varepsilon$ , is the Euler constant ( $\varepsilon = 0.5772$ ). The Gumbel law has a constant skewness coefficient of 1.1396 (Gumbel [8]).



To introduce the sorting operator in the statistical analysis based on the Gumbel law, the least-square method, LSM, must be apply to the parameters estimation, according to the following equations (Henriques [9]):

$$\hat{\alpha} = \frac{n \sum_{k=1}^n x_k^2 - \left( \sum_{k=1}^n x_k \right)^2}{n \sum_{k=1}^n z_k x_k - \sum_{k=1}^n x_k \sum_{k=1}^n z_k} \quad \hat{u} = \bar{x} + \hat{\alpha} \frac{\sum_{k=1}^n z_k}{n} \quad (5) \text{ and } (6)$$

where  $\hat{\alpha}$  and  $\hat{u}$  are the estimates of  $\alpha$  and  $u$  respectively and  $z_k$  depends upon the empirical non-exceedance probability  $F(x_k)$  of the element  $x_k$  of the sample:

$$z_k = -\ln(-\ln(F(x_k))) \quad (7)$$

In the comparison carried out among plotting positions formulas based or not on the sorting operator, SO, the Gringorten formula was applied as this formula is generally recommended when a statistical law for extremes, like the Gumbel law, is under consideration (Gringorten [7], Stedinger *et al.* [16] and Cunnane [2]):

$$F(x_k) = \frac{i - 0.44}{n + 0.12} \quad (8)$$

### 3 Sorting operator, SO

Let us consider two watersheds located in the same homogenous hydrologic region and geographically near such that the same extreme weather conditions affect both watersheds causing extreme rainfall events and therefore extreme floods in the same or in vicinal periods (temporal proximity criterion) and with similar relative magnitudes (relative rank proximity criterion).

Let  $y_1, y_2, \dots, y_i, \dots, y_k$  and  $x_1, x_2, \dots, x_i, \dots, x_n$  represent the samples of the annual maximum instantaneous discharges,  $Q_{ami}$ , in the two watersheds, the lengths  $n_y$  and  $n_x$  being such that  $n_y \ll n_x$ . Those samples are further denoted by “censored” and “complete” samples, respectively. According to the SO concept, to assign an empirical non-exceedance probability to each  $y_i$  value of the censored sample, the complete sample is firstly sorted (in ascending order). The year in which each  $x_i$  value occurred and the rank of such value in the complete sample are stored in the sorting operator, SO. By applying a plotting position formula, the empirical probability of  $y_i$  is then computed based on the rank of the discharge of the complete sample that according to the time proximity criterion corresponds to  $y_i$ . In other words, the empirical non-exceedance probability of  $y_i$  is computed based on the rank stored in the SO for the year in which  $y_i$  occurred (Delgado [4] and Portela and Delgado [12]).

Table 1 exemplifies how to apply the sorting operator, SO. The right side of the table contains the complete and the censored samples and the left side the



Table 1: Samples of Qami in watersheds 1,  $x_i$ , and 2,  $y_i$  and sorting operator, SO, utilized to assign non-exceedance probabilities to the  $y_i$  values.

Hydrologic year	Annual maximum instantaneous discharges, Qami		Hydrologic year	Complete sample: ranking of the annual maximum instantaneous discharges, Qami			Sorting Operator SO
	Watershed 1, $x_i$ (complete sample) (m <sup>3</sup> /s)	Watershed 2, $y_i$ (censored sample) (m <sup>3</sup> /s)		Rank	Discharge		
					Watershed 1, $x_i$ (m <sup>3</sup> /s)	Watershed 2, $y_i$ (m <sup>3</sup> /s)	
1966/67	511	475	1982/83	1	187	--	--
1967/68	688	--	1983/84	2	260	--	--
1968/69	882	602	1990/91	3	310	296	(1990/91;3)
1969/70	--	--	1977/78	4	395	--	--
1970/71	558	492	1981/82	5	421	374	(1981/82;5)
1971/72	1933	1144	1974/75	6	433	376	(1974/75;6)
1972/73	1518	--	1984/85	7	477	--	--
1973/74	1035	612	1988/89	8	499	--	--
1974/75	433	376	1966/67	9	511	475	(1966/67;9)
1975/76	--	--	1970/71	10	558	492	(1970/71;10)
1976/77	--	--	1978/79	11	635	548	(1978/79;11)
1977/78	395	--	1967/68	12	688	--	--
1978/79	635	548	1987/88	13	850	--	--
1979/80	1769	--	1968/69	14	882	602	(1968/69;14)
1980/81	1210	--	1985/86	15	990	607	(1985/86;15)
1981/82	421	374	1973/74	16	1035	612	(1973/74;16)
1982/83	187	--	1980/81	17	1210	--	--
1983/84	260	--	1972/73	18	1518	--	--
1984/85	477	--	1979/80	19	1769	--	--
1985/86	990	607	1971/72	20	1933	1144	(1971/72;20)
1986/87	--	--	1969/70	--	--	--	--
1987/88	850	--	1975/76	--	--	--	--
1988/89	499	--	1976/77	--	--	--	--
1989/90	--	--	1986/87	--	--	--	--
1990/91	310	296	1989/90	--	--	--	--

Gringorten formula					
Without sorting operator (without SO)			With sorting operator (with SO)		
Rank	$y_i$ (m <sup>3</sup> /s)	F( $y_i$ ) (m <sup>3</sup> /s)	Rank	$y_i$ (m <sup>3</sup> /s)	F( $y_i$ ) (m <sup>3</sup> /s)
1	296	0.0553	3	296	0.1272
2	374	0.1542	5	374	0.2266
3	376	0.2530	6	376	0.2763
4	475	0.3518	9	475	0.4254
5	492	0.4506	10	492	0.4751
6	548	0.5494	11	548	0.5249
7	602	0.6482	12	602	0.5746
8	607	0.7470	15	607	0.7237
9	612	0.8458	16	612	0.7734
10	1144	0.9447	20	1144	0.9722

Parameters of the Gumbel law given by the least-square method			
$\alpha$	212.77	$\alpha$	193.79
u	437.49	u	426.52

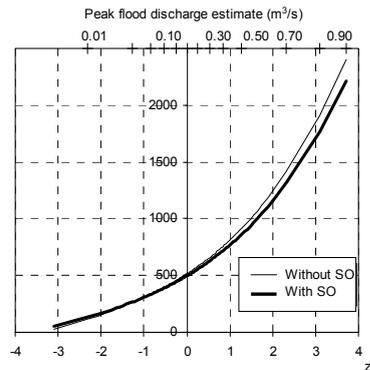


Figure 3: Statistical analysis of the censored sample without and with considering the sorting operator: empirical non-exceedance probabilities and parameters and probability distribution function of the Gumbel law.

ranking of the former sample as well as the sorting operator and the assignment among values of the SO and of the censored sample.

In Figure 3 a comparison between the application to the censored sample of the Gumbel law based or not in the SO is accomplished. The table included in

such figure contains the empirical non-exceedance probabilities evaluated by the Gringorten plotting position formula without and with considering the ranks given by the SO as well as the respective estimates by the least-square method of the parameters of the Gumbel law. The diagram represents the Gumbel probability distribution function also without and with considering the sorting operator. It should be stressed that when the SO is not considered the probability plotting formula accounts for the rank of each  $y_i$  element in the censored sample alone, as shown in the table of Figure 3.

Figure 3 clearly shows the effect of SO in the shape of the Gumbel probability distribution function and, therefore, in the estimates of peak flood discharges. However, before recommending the application of the SO, it is fundamental to analyze if it enables or not more accurate estimates.

#### 4 Testing the method

To test the SO-based model, applications to samples of Qami in several Portuguese gage stations were carried out. For each one of those samples – complete sample with length  $n = 100$  different random censored samples with length  $n_j$  ( $n_j < n$ ) were built. Lengths of 5, 10 and 20 years were adopted for  $n_j$ . To facilitate the generation process of the 100 censored samples only stations with more than 25 to 30 years of Qami records were considered. In fact, the number of different censored samples that can be built based on a sample of length  $n$  is given by

$$\binom{n}{n_j} = \frac{n!}{(n - n_j)! n_j!} \quad (9)$$

which means that a complete sample of 25-years of records comprehends 53130 different censored samples with length  $n_j=20$  which facilitates the process of picking up 100 random censored samples. For smaller values of  $n_j$  or for larger values of  $n$  the number of censored samples is bigger. The number of 100 censored samples with a given length was considered to be big enough to guarantee the randomness of the distribution of the estimates of peak flood discharges based on those samples and, at the same time, small enough to allow the analysis even in those basins with fewer records.

For each complete sample with length  $n$  the generation of the 100 different censored samples each one with a given length  $n_j$  was developed according to: a) by applying combinatory analysis, definition of all the possible censored samples with length  $n_j$ ; b) assignment of the sequence of those samples to a 0-1 transformed scale; c) generation of 100 uniformly distributed random numbers; d) collection of the 100 censored samples with sequence numbers closer to the 100 generated random numbers.

For each set of 100 censored samples with length  $n_j$ , peak flood discharges with different return periods,  $T$  (the relationship between  $F(x)$  and return period  $T$  is given by  $F(x) = 1 - 1/T$ ), were estimated by applying the Gumbel law with

parameters given by the LSM with empirical non-exceedance probabilities computed by the Grigorten formula without and with considering the sorting operator. The procedures and the results achieved based on the least-square method combined with the sorting operator will be further denoted by LSM +  $\Omega$  or simply by  $\Omega$ , and those obtained without considering the sorting operator by LSM +  $\Gamma$  or simply by  $\Gamma$ . The estimates, for a given return period, T, of the peak flood discharges obtained considering the SO or without considering such operator will be denoted by  $\hat{Q}_{\Omega,i}^{nj}(T)$  and  $\hat{Q}_{\Gamma,i}^{nj}(T)$ , respectively, where the upper index nj is fixed and i varies between 1 and 100.

The peak flood discharges estimated based on a given set of 100 censored samples with the fixed length nj were characterized by their average and standard deviation, according to the following equations:

- Average without –  $\hat{Q}_{\Gamma}^{nj}(T)$  – and with –  $\hat{Q}_{\Omega}^{nj}(T)$  – the SO:

$$\hat{Q}_{\Gamma}^{nj}(T) = \frac{\sum_{i=1}^{100} \hat{Q}_{\Gamma,i}^{nj}(T)}{100} \quad \hat{Q}_{\Omega}^{nj}(T) = \frac{\sum_{i=1}^{100} \hat{Q}_{\Omega,i}^{nj}(T)}{100} \quad (10) \text{ and } (11)$$

- Standard-deviation without –  $\sigma_{\Gamma}^{nj}(T)$  – and with –  $\sigma_{\Omega}^{nj}(T)$  – the SO:

$$\sigma_{\Gamma}^{nj}(T) = \sqrt{\frac{\sum_{i=1}^{100} [\hat{Q}_{\Gamma,i}^{nj}(T) - \hat{Q}_{\Gamma}^{nj}(T)]^2}{99}} \quad \sigma_{\Omega}^{nj}(T) = \sqrt{\frac{\sum_{i=1}^{100} [\hat{Q}_{\Omega,i}^{nj}(T) - \hat{Q}_{\Omega}^{nj}(T)]^2}{99}} \quad (12) \text{ and } (13)$$

Along with the previous computations the peak flood discharge provided by the complete sample for each T was also estimated,  $\hat{Q}(T)$ . Regarding this estimates it is pointless to speak about the SO as the whole sample is utilized to compute the empirical probability as well as the parameters of the Gumbel law.

The relative performance of the discharges obtained for a given return period, T, by applying the LSM +  $\Gamma$  and LSM +  $\Omega$  models to the 100 censored samples were assessed in terms of the bias –  $BIAS_{\Gamma}^{nj}(T)$  and  $BIAS_{\Omega}^{nj}(T)$  – and of the standard deviation –  $STD_{\Gamma}^{nj}(T)$  and  $STD_{\Omega}^{nj}(T)$  –, both measures made non-dimensional by division by the peak flood discharge computed based on the complete sample,  $\hat{Q}(T)$ , according to:

- Bias without –  $BIAS_{\Gamma}^{nj}(T)/\hat{Q}(T)$  – and with –  $BIAS_{\Omega}^{nj}(T)/\hat{Q}(T)$  – the SO:

$$\frac{BIAS_{\Gamma}^{nj}(T)}{\hat{Q}(T)} = \frac{\hat{Q}(T) - \hat{Q}_{\Gamma}^{nj}(T)}{\hat{Q}(T)} = 1 - \frac{\hat{Q}_{\Gamma}^{nj}(T)}{\hat{Q}(T)} \quad (14)$$

$$\frac{BIAS_{\Omega}^{nj}(T)}{\hat{Q}(T)} = \frac{\hat{Q}(T) - \hat{Q}_{\Omega}^{nj}(T)}{\hat{Q}(T)} = 1 - \frac{\hat{Q}_{\Omega}^{nj}(T)}{\hat{Q}(T)} \quad (15)$$



- Non-dimensional standard-deviation without –  $STD_{\Gamma}^{nj}(T)/\widehat{Q}(T)$  – and with –  $STD_{\Omega}^{nj}(T)/\widehat{Q}(T)$  – the SO:

$$\frac{STD_{\Gamma}^{nj}(T)}{\widehat{Q}(T)} = \frac{\sigma_{\Gamma}^{nj}(T)}{\widehat{Q}(T)} \quad \frac{STD_{\Omega}^{nj}(T)}{\widehat{Q}(T)} = \frac{\sigma_{\Omega}^{nj}(T)}{\widehat{Q}(T)} \quad (16) \text{ and } (17)$$

The relative efficiency of the estimates given by the  $LSM + \Gamma$  and the  $LSM + \Omega$  models was also assessed by the ratio between the mean square errors, MSE, of such estimates:

$$\frac{MSE_{\Omega}^{nj}(T)}{MSE_{\Gamma}^{nj}(T)} = \frac{[\sigma_{\Omega}^{nj}(T)]^2 + [BIAS_{\Omega}^{nj}(T)]^2}{[\sigma_{\Gamma}^{nj}(T)]^2 + [BIAS_{\Gamma}^{nj}(T)]^2} \quad (18)$$

## 5 Validating the method

Having concluded, from the testing phase, that the SO-based models resulted in better estimates of the peak flood discharge, a validation phase was carried out supported by several groups of two neighboring watersheds located in the same homogenous hydrologic regions and having Qami records.

Let  $Q1_i$  and  $Q2_i$  denote two samples of Qami in watersheds 1 (length  $n1$ ) and 2 (length  $n2$ ), respectively, and let assume that by applying the temporal proximity criterion it is possible to identify the pairs of values  $Q1_i$  and  $Q2_i$  that correspond to the same flood events. To validate the SO-based approaches a procedure equivalent to the one described for the testing phase was applied, according to the following main steps in which the sample of watershed 2 is used only to define the sorting operator that is utilized to categorize the values of the sample in watershed 1: i) based on the  $Q2_i$  sample, definition of the SO; ii) based on the  $Q1_i$  sample, definition of 100 random censored samples with length  $n_j$  such as  $n_j < n1$ ; iii) ranking of the elements of each censored sample based on the sorting operator –  $\Omega$  models, more probably with discontinuous numbering between 1 and  $n2$  – or taking into account the censored sample alone –  $\Gamma$  models with sequential numbering between 1 and  $n_j$ ; iv) for basin 1, estimation of peak flood discharges by applying the Gumbel law with parameters given by LSQ based on the Grigorten formula applied taking into account or not the sorting operator – models  $LSM + \Omega$  and  $LSM + \Gamma$ , respectively; and v) by applying the performance measures presented in the testing phase, comparison, between peak flood discharges provided by the censored samples with and without the SO and by the complete  $Q1_i$  sample.

## 6 Results

The watersheds adopted as case studies are specified in Table 2. The horizontal dashed lines identify the sets of two watersheds considered in the validation phase. The watersheds of each set belong to a same homogenous hydrologic



region according to the homogenous region mapping developed by Dias [5] and Portela and Dias [13] and are geographically close.

For the testing phase, Table 3 exemplifies the results achieved based on 100 censored sample with lengths,  $n_j$ , of 5, 10 and 20 years built upon the complete samples of Qami in Vinhais-Quinta da Ranca and in Fragas da Torre.

To stress the relevance of the SO, namely when very small samples are under consideration, the estimates of the 100 and 500-years peak flood discharges based on censored samples with  $n_j = 5$  without and with considering the SO are systematized in Table 4 for all the case studies.

Table 2: Watersheds adopted as case studies. General features.

Stream gage	River	Watershed area (km <sup>2</sup> )	Length of the Qami sample
Rebordelo	Rabaçal	868.62	34
Vinhais-Quinta da Ranca	Tuela	478.55	29
Vale Trevo	Masueime	405.40	38
Cidadelehe	Côa	1743.10	48
Fragas da Torre	Paiva	646.75	49
Castro Daire	Paiva	288.17	49
Pte de Canavezes	Tâmega	3163.33	27
Cabriz	Sampaio	10.80	21

Table 3: Testing phase. Stream gages of Vinhais-Quinta da Ranca and Fragas da Torre. Censored samples with 5, 10 and 20 years. Return periods, T, of 100, 250, 500 and 1000 years.

n <sub>j</sub>	T	Vinhais-Quinta da Ranca									Fragas da Torre								
		Q̂(T)	Q̂ <sup>Ω</sup> (T)	Q̂ <sup>Γ</sup> (T)	BIAS <sup>Ω</sup> <sub>1</sub> (T)	BIAS <sup>Γ</sup> <sub>1</sub> (T)	STD <sup>Ω</sup> <sub>1</sub> (T)	STD <sup>Γ</sup> <sub>1</sub> (T)	MSE <sup>Ω</sup> <sub>1</sub> (T)	MSE <sup>Γ</sup> <sub>1</sub> (T)	Q̂(T)	Q̂ <sup>Ω</sup> (T)	Q̂ <sup>Γ</sup> (T)	BIAS <sup>Ω</sup> <sub>2</sub> (T)	BIAS <sup>Γ</sup> <sub>2</sub> (T)	STD <sup>Ω</sup> <sub>2</sub> (T)	STD <sup>Γ</sup> <sub>2</sub> (T)	MSE <sup>Ω</sup> <sub>2</sub> (T)	MSE <sup>Γ</sup> <sub>2</sub> (T)
		(m <sup>3</sup> /s)	(m <sup>3</sup> /s)	(m <sup>3</sup> /s)	(-)	(-)	(-)	(-)	(-)	(-)	(m <sup>3</sup> /s)	(m <sup>3</sup> /s)	(m <sup>3</sup> /s)	(-)	(-)	(-)	(-)	(-)	(-)
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[18]	
(year)	(year)	(m <sup>3</sup> /s)	(m <sup>3</sup> /s)	(m <sup>3</sup> /s)	(-)	(-)	(-)	(-)	(-)	(m <sup>3</sup> /s)	(m <sup>3</sup> /s)	(m <sup>3</sup> /s)	(-)	(-)	(-)	(-)	(-)	(-)	
5	100	693	646	649	-0.068	-0.063	0.379	0.104	0.10	1075	1036	1119	-0.037	0.041	0.300	0.113	0.16		
	250	796	738	743	-0.072	-0.067	0.392	0.109	0.10	1214	1163	1265	-0.042	0.042	0.318	0.121	0.16		
	500	874	808	814	-0.075	-0.069	0.400	0.112	0.10	1319	1260	1376	-0.045	0.043	0.329	0.126	0.16		
	1000	951	877	884	-0.078	-0.070	0.407	0.114	0.11	1423	1356	1486	-0.047	0.044	0.338	0.130	0.16		
10	100	693	666	667	-0.039	-0.037	0.237	0.088	0.16	1075	1013	1110	-0.058	0.032	0.234	0.096	0.18		
	250	796	763	765	-0.042	-0.039	0.246	0.092	0.16	1214	1136	1255	-0.064	0.034	0.249	0.101	0.17		
	500	874	836	838	-0.043	-0.040	0.251	0.095	0.17	1319	1229	1365	-0.068	0.035	0.259	0.105	0.17		
	1000	951	909	912	-0.045	-0.041	0.255	0.097	0.17	1423	1322	1474	-0.071	0.036	0.267	0.108	0.17		
20	100	693	656	671	-0.054	-0.031	0.127	0.065	0.27	1075	1078	1066	0.002	-0.009	0.106	0.038	0.13		
	250	796	751	770	-0.056	-0.033	0.132	0.069	0.28	1214	1216	1202	0.002	-0.010	0.113	0.040	0.13		
	500	874	824	844	-0.057	-0.033	0.135	0.071	0.28	1319	1320	1305	0.001	-0.010	0.116	0.042	0.14		
	1000	951	896	919	-0.058	-0.034	0.138	0.073	0.29	1423	1425	1408	0.001	-0.011	0.120	0.043	0.14		

Tables 3 and 4 clearly show that the SO improves significantly the estimates of the peak flood discharges, as denoted by the comparison between LSM + Ω and LSM + Γ models of the non-dimensional values of the bias and of the standard deviation and by the very small values of the mean square errors.

It should be stressed that each discharge of columns 4, 5, 12 and 13 of Table 3 and of columns 4 and 5 of Table 4 is an average of 100 estimates, each estimate provided by a censored sample with length  $n_j$ , according to equations (10) and (11) which means that in terms of the appraisal of the performance of the models the standard deviation is as much relevant as the average. In fact, the standard deviation provides a measure of the dispersion around the average and therefore



a high deviation would mean that the estimates could be rather similar to  $\hat{Q}(T)$  or otherwise quite different from  $\hat{Q}(T)$ . Smaller standard deviations, as those obtained by the SO-based models, mean that the estimates provided by the censored sample are consistently closer to  $\hat{Q}(T)$ .

Table 4: Testing phase. 5-years censored samples. Return periods, T, of 100 and 500 years.

Stream gage [1]	T [2] (year)	$\hat{Q}(T)$ [3] (m <sup>3</sup> /s)	$\hat{Q}_{\Gamma}^n(T)$ [4] (m <sup>3</sup> /s)	$\hat{Q}_{\Omega}^n(T)$ [5] (m <sup>3</sup> /s)	BIAS $\frac{n}{T}$ (T)	BIAS $\frac{n}{\Omega}$ (T)	STD $\frac{n}{T}$ (T)	STD $\frac{n}{\Omega}$ (T)	MSE $\frac{n}{T}$ (T)
					$\hat{Q}(T)$ [6] (-)	$\hat{Q}(T)$ [7] (-)	$\hat{Q}(T)$ [8] (-)	$\hat{Q}(T)$ [9] (-)	MSE $\frac{n}{T}$ (T) [10] (-)
Rebordelo	100	1199	1119	1275	-0.067	0.063	0.253	0.141	0.350
	500	1525	1415	1630	-0.072	0.069	0.253	0.150	0.400
Vinhais-Qta Ranca	100	693	646	649	-0.068	-0.063	0.379	0.104	0.100
	500	874	808	814	-0.075	-0.069	0.400	0.112	0.100
Vale Trevo	100	193	214	186	-0.109	-0.033	0.346	0.110	0.100
	500	247	275	239	-0.113	-0.035	0.356	0.118	0.110
Cidadelehe	100	1057	1072	1029	-0.015	-0.026	0.365	0.117	0.110
	500	1358	1376	1322	-0.013	-0.027	0.380	0.123	0.110
Fragas da Torre	100	1075	1036	1119	-0.037	0.041	0.300	0.113	0.160
	500	1319	1260	1376	-0.045	0.043	0.329	0.126	0.160
Castro Daire	100	430	368	457	-0.145	0.064	0.260	0.137	0.260
	500	531	443	566	-0.166	0.065	0.287	0.153	0.250
Ponte de Canavezes	100	2542	2874	2635	0.131	0.037	0.124	0.096	0.320
	500	3142	3595	3267	0.144	0.040	0.133	0.107	0.340
Cabriz	100	33	32	33	-0.015	0.023	0.284	0.140	0.250
	500	41	40	42	-0.019	0.023	0.305	0.152	0.250

The results from the validation phase are exemplified in Table 5 for the stream gages of Fragas da Torre and Castro Daire. Only lengths of the censored samples of 5 and 10 years are exemplified as the SO is particularly advantageous when small samples need to be considered. Also for larger censored samples the results become always more accurate.

Columns 3 to 10 of Table 5 relate to estimates of peak flood discharges at Fragas da Torre based on two sets of 100-censored samples each, one with  $n_j = 5$  and the other with  $n_j = 10$ . Each censored sample was ranked taking into account only its own elements or by applying the sorting operator – results denoted by  $\Gamma$  and  $\Omega$ , respectively. The SO was defined based on the complete sample of Castro Daire. In the SO-based results, the ranking of each censored sample was given by the values stored in the SO operator according to the temporal proximity of the floods events at Fragas da Torre and Castro Daire watersheds, as exemplified in Table 1. Columns 11 and 18 have similar meaning for Castro Daire the respective SO operator having been defined based on the complete sample of Fragas da Torre. Once more, the Grigorten formula was applied to evaluate the empirical non-exceedance probabilities.

Table 5 shows that for Fragas da Torre the SO-based peak flood discharge estimates are always closer to  $\hat{Q}(T)$  (more accurate estimates) while in Castro Daire this only happens for  $n_j = 10$ : in fact the estimates based on the 5-years censored samples without SO are better (closer to  $\hat{Q}(T)$ ) than those provided by the SO.

Table 5: Validation phase. 5 and 10-years censored samples. Top: flood discharges at Fragas da Torre for SO provided by the complete Qami sample at Castro Daire. Down: flood discharges at Castro Daire for SO provided by the complete Qami sample at Fragas da Torre.

Peak flood discharges at Fragas da Torre. Sorting operator, SO, defined based on the complete sample of annual maximum instantaneous discharges, Qami, at Castro Daire										
nj	T	$\bar{Q}(T)$	$\hat{Q}_F^{nj}(T)$	$\hat{Q}_\Omega^{nj}(T)$	$\text{BIAS}_F^{nj}(T)$	$\text{BIAS}_\Omega^{nj}(T)$	$\text{STD}_F^{nj}(T)$	$\text{STD}_\Omega^{nj}(T)$	$\text{MSE}_\Omega^{nj}(T)$	$\text{MSE}_F^{nj}(T)$
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[10]
(anos)	(anos)	(m <sup>3</sup> /s)	(m <sup>3</sup> /s)	(m <sup>3</sup> /s)	(-)	(-)	(-)	(-)	(-)	(-)
5	100	1075	1036	1104	0.037	-0.027	0.300	0.288	0.92	
	500	1319	1260	1353	0.045	-0.026	0.329	0.320	0.94	
	1000	1423	1356	1460	0.047	-0.026	0.338	0.330	0.94	
	5000	1666	1579	1709	0.052	-0.026	0.356	0.349	0.95	
	100	1075	1013	1055	0.058	0.019	0.234	0.170	0.50	
10	500	1319	1229	1287	0.068	0.024	0.259	0.189	0.51	
	1000	1423	1322	1387	0.071	0.026	0.267	0.196	0.51	
	5000	1666	1537	1618	0.077	0.029	0.283	0.208	0.51	
	100	1075	1013	1055	0.058	0.019	0.234	0.170	0.50	
	500	1319	1229	1287	0.068	0.024	0.259	0.189	0.51	

Peak flood discharges at Castro Daire. Sorting operator, SO, defined based on the complete sample of annual maximum instantaneous discharges, Qami, at Fragas da Torre										
nj	T	$\bar{Q}(T)$	$\hat{Q}_F^{nj}(T)$	$\hat{Q}_\Omega^{nj}(T)$	$\text{BIAS}_F^{nj}(T)$	$\text{BIAS}_\Omega^{nj}(T)$	$\text{STD}_F^{nj}(T)$	$\text{STD}_\Omega^{nj}(T)$	$\text{MSE}_\Omega^{nj}(T)$	$\text{MSE}_F^{nj}(T)$
[1]	[2]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[18]
(anos)	(anos)	(m <sup>3</sup> /s)	(m <sup>3</sup> /s)	(m <sup>3</sup> /s)	(-)	(-)	(-)	(-)	(-)	(-)
5	100	418	408	436	0.023	-0.043	0.238	0.326	1.90	
	500	513	498	534	0.028	-0.041	0.265	0.361	1.87	
	1000	554	537	576	0.030	-0.041	0.274	0.373	1.85	
	5000	648	627	674	0.033	-0.040	0.292	0.394	1.82	
	100	418	395	416	0.055	0.004	0.206	0.194	0.82	
10	500	513	480	509	0.064	0.007	0.230	0.212	0.79	
	1000	554	517	549	0.067	0.008	0.238	0.218	0.78	
	5000	648	601	642	0.072	0.010	0.253	0.229	0.76	
	100	418	395	416	0.055	0.004	0.206	0.194	0.82	
	500	513	480	509	0.064	0.007	0.230	0.212	0.79	

Although, for the different case studies, the better estimates were much more often provided by the SO, the results from the validation phase were not as straightforward as those from the testing phase most probably because additional criteria is still need to identify the stream gage that should be utilized to define the sorting operator. In fact, the assumption of watersheds located in the same homogenous hydrologic region and geographically close may not be enough to ensure that the flood events registered in the same or in a nearby period will have similar relative magnitude.

## 7 Final remarks

A model aiming at improving the estimates of the peak flood discharges specially when based on small samples was developed. The model acts in the definition of the empirical non-exceedance probability by means of a sorting operator, SO, and considers the application of the Gumbel law with parameters evaluated according to the least square method, LSM.



The applications carried out showed that despite its great potential the method still requires improvement namely regarding the development of criteria aiming at identifying pairs of watershed in which it is expectable to have peak flood discharges with similar relative magnitudes occurring in the same period or in vicinal periods. Also other laws of extreme besides the Gumbel law and compatible with the parameter estimation by the least square method – like the gama law – should be tested.

## References

- [1] Coursey, D. G. D., 1973. “Objective regionalization of peak flow rates”, in E. F. Schulz, V. A. Koelzer & K. Mahmood, eds, *Floods and Droughts*, Water Resources Publications, pp. 395-405.
- [2] Cunnane, C., 1978, “Unbiased plotting positions – a review”, *Journal of Hydrology* (37), pp. 205-222.
- [3] Darlymple, T., 1960. “Flood frequency analysis”. U.S. Geological Survey *Water Supply Paper* 1543-A., Reston, Va.
- [4] Delgado, J. M. M., 2007. *Flood discharge evaluation under scarcity of data*, MSc Thesis in Civil Engineering, IST, Lisbon.
- [5] Dias, A. T., 2003. *Caudais instantâneos máximos anuais em Portugal Continental. Proposta de regionalização* (in Portuguese). MSc Thesis in Hydraulics and Water Resources, IST, Lisbon.
- [6] Gonçalves, M. I., 1993. *Regionalização da distribuição generalizada de extremos. Aplicação à previsão de caudais de cheia*. (in Portuguese). MSc Thesis in Hydraulics and Water Resources, IST, Lisbon.
- [7] Gringorten, I. I., 1963. “A plotting rule for extreme probability paper”, *Journal of Geophysical Research* 68 (3), pp. 813–814.
- [8] Gumbel, E. J., 1958. *Statistics of extremes*. Columbia University Press, New York.
- [9] Henriques, A. N. F. G., 1990. *Modelos de distribuição de Frequências de Caudais de Cheia* (in Portuguese), PhD Thesis, IST, Lisbon.
- [10] Lettenmaier, D. P.; Wallis, J. R. & Wood, E. F., 1987. “Effect of regional heterogeneity on flood frequency estimation”. *Water Resources Research*, 23 (2), pp. 313-323.
- [11] Loureiro, J. M., 1984. “Expressão para o cálculo do caudal de cheia em cursos de água de Portugal”, *Recursos Hídricos*, 5 (1), pp. 53-78.
- [12] Portela, M. M.; Delgado, J. M., 2008. “Estimação de caudais de ponta de cheia em situação de escassez de registos de caudais instantâneos máximos anuais”. *9º Congresso da Água*, Associação Portuguesa dos Recursos Hídricos, APRH, Cascais.
- [13] Portela, M. M.; Silva, A. T., 2005, “Application of the index-flood method to the regionalization of flood peak discharges in Portugal mainland”, *River Basin Management III* (Bologna, Italy), Wessex Institute, WITPress, ISBN 1-84564-023-3, pp. 475-485, Southampton, Boston.
- [14] Raynall, J. A. & Salas, J. D., 1986. “Estimation procedures for the type-1 extreme value distribution”, *Journal of Hydrology* (87), 315-336.



- [15] Reiss, R.-D. & Thomas, M., 2001. *Statistical analysis of extreme values*, Birkhauser Verlag.
- [16] Stedinger, J. R.; Vogel, R. M.; Foufoula-Georgiou, E, 1992. "Frequency analysis of extreme events", in *Handbook of Hydrology*, chapter 18, edited by D.A. Maidment, McGraw-Hill, New York.

