ASSESSMENT OF EXTREME FLOOD PRODUCTION MECHANISMS THROUGH POT ANALYSIS OF DAILY DATA

Pierluigi Claps [1], Francesco Laio [1] and Paolo Villani [2]

[1] Dipartimento di Idraulica, Trasporti e Infrastrutture Civili, Politecnico di Torino, C.so Duca degli Abruzzi 24, 10129 Torino, Italy [claps@polito.it; laio@polito.it]

[2] Dipartimento di Ingegneria Civile – Università di Salerno, Via Ponte Don Melillo, 84084 Fisciano (SA), Italy [villani@bridge.diima.unisa.it]

SUMMARY

As an attempt to improve flood risk prediction we investigate the relations between the stochastic processes of daily rainfall and runoff. We consider a simple stochastic model of the extreme precipitation in which the occurrences are mutually independent and the distribution of marks can be derived through the analysis of the Peak Over Threshold (POT) process. Our initial interest concentrates on the relationships between the mean annual number of rainfall and runoff peak events (Λ_p and Λ_q) at various thresholds: the mechanism of flood production and its changes are investigated by analysing the nature of these relationships. To this aim we consider that the watershed operates as a stochastic filter on the rainfall process, producing a filtered point process that reproduces occurrence and magnitude of floods. The analysis of the filter mechanism provides the relationships between stochastic parameters of maximum rainfall and discharge. Particularly, we consider a simple filter wherein the runoff peaks are produced after a volume abstraction on precipitation, which essentially limits the number of runoff events. Since Λ_p and Λ_q are parameters playing a significant role in Poisson-derived flood frequency distributions, we investigate the possible dependence of the obtained results on the hypothesis of Poissonian occurrences and on the distribution of marks,

We derived relations between the frequency of rainfall and flood events at varying thresholds for several observed series of rainfall and runoff. The different patterns of the basin transformation found seem to confirm the stochastic nature of the basin filter and its strong relations with the hydro-geologic and climatic characteristics of the basin.

Keywords: Floods, Rainfall, Extreme Events, POT Analysis.

1. INTRODUCTION

The research and application aspects related to regional flood frequency analysis have reached very good standards, but several controversial points still prevent the agreement on universally recognised procedures (see e.g. Bobée et al., 1993). Recent approaches agree on the necessity to maximise geomorphoclimatic information behind the flood formation mechanisms, either through models based on derived distributions (e.g. Iacobellis and Fiorentino, 2000) or by derivation of the flood curve by simulation (e.g. Hashemi et al., 2000).

Less attention has been given to the possibility of a direct statistical analysis of continuous rainfall and runoff time series, which can provide greater detail on the transformation mechanisms with respect to annual maximum series. In fact, additional information deriving from continuous daily records improves the knowledge of the base rainfall process and increases the number of observed flood peaks. This procedure therefore augments the possibilities of a data-based assessment of the hypotheses underlying the definition of a flood frequency distribution within a geomorphoclimatic framework.

The mentioned analysis can be carried out through the application of Peak Over Threshold (POT) techniques on daily rainfall and runoff records. Lang et al. (1999) recently pointed out pros and cons of POT-based flood frequency estimation. Main drawback of the technique is the difficulty to set objective criteria for the choice of the threshold values which, on the other hand, is a crucial point of the procedure.

In this paper, analysis of continuous data is presented with a POT approach that allows one to naturally and objectively consider the threshold values. To this end, relations between the number of rainfall and runoff event over a given threshold have been linked to the magnitude of the events

themselves, investigating the attenuation provided by the basin over the entire range of runoff peak values.
This data-based approach utilises statistical relations between rainfall and runoff peaks to explore the interaction between the empirical distributions of the two variables in different classes of drainage basins. The envisaged statistical structure is compatible with the estimation of parameters related to the geomorphoclimatic configuration of the basin and the consequent transfer of hydrologic information towards ungauged basins.

2. POT ANALYSIS

2.1 Procedures of peak selection

For a given catchment we consider the time series of daily rainfall and discharge: since both variables vary randomly in time, they can be represented as continuous stochastic processes (sampled at the daily time scale). The first step towards an analysis of the relationship between extreme rain and flood events is to select, from the continuous time series, those values that can reasonably be considered as peak events. The usual procedure for peaks selection is to retain only those peaks that exceed a certain threshold value s_0, with a procedure known as Peak Over Threshold, or POT (e.g., Todorovic, 1978; Madsen et al. ,1997; Lang et al., 1999). Each peak event is considered to last for the whole time period the discharge (or rainfall) remains above the threshold (see Figure 2-1).
This procedure of peaks selection presents some well-known difficulties in the definition of the base level s_0 (e.g., Lang et al., 1999; Robson and Reed, 1999). With low s_0 values nearly the whole time series is above the base level and only very few peaks are identified (in fact, if the discharge never downcrosses s_0 a unique peak is found) while with large s_0 only few very high peaks are retained. To have available a significant number of peaks, s_0 should be taken at intermediate values, but this arises problems of lack of mutual independence among the peaks (that is a prerequisite to any statistical frequency analysis). In the usual POT procedure it is often imposed that a peak is retained only if it is separated from the previous one by a given number of days and if the discharge before the event drops well below s_0. In addition, actual use of these criteria varies from one case to the other, with the result that a degree of subjectivity is added to the procedure.

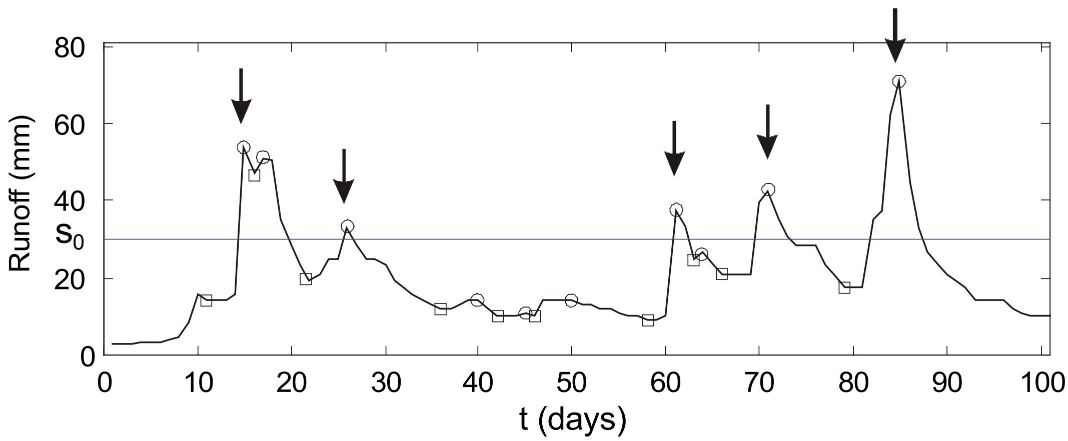

Figure 2-1: Sketch of the procedure for peaks selection from the time series of daily runoff. Arrows represent the peaks selected with the usual POT procedure; open circles and squares are the local maxima and minima used in the alternative procedure for peaks selection.

To limit the subjective characters of the procedure, we decided to adopt an alternative two-steps method: (i) the events are identified in correspondence to all the local maxima of the time series (open circles in Figure 2-1); (ii) the magnitude of each event is then assigned in two alternative ways: (a) the actual ordinate of the maximum (method M1, gray bars in Figure 2-2), or (b) the difference between the maximum and the first minimum preceding the event (method M2, black bars in Figure 2-2). The first choice has the advantage that the actual value of the discharge is preserved, while with the second method the mutual dependence of subsequent peaks is reduced. For example, consider two peaks very close in time (e.g. the first and the second in Figure 2-2): the magnitude of the second peak can result artificially increased by a value that depends on the intensity of the previous event, which has not yet exhausted its effects on the basin. Alternatively, if the ordinate of the minimum were subtracted, the mutual dependence would be eliminated (however, other sources of correlation, due for example to rainfall persistency, can still remain). As such, M2 events do not strictly represent the real discharge; rather, they approximate the effective rainfall component, i.e. the fraction of rainfall that becomes runoff within one day. Both M1 and M2 methods will be employed in the following of the paper.

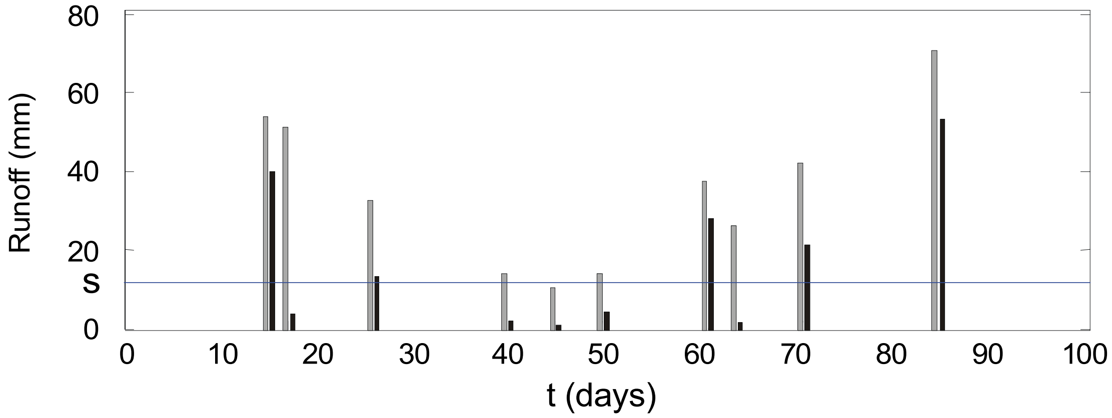

Figure 2-2. Marked point process obtained after peak selection from the discharge time series shown in Figure 2-1. Gray bars: method M1; black bars: method M2.

2.2 Basic statistical relationships and their conditions of validity

Whatever procedure of peak selection is chosen, the continuous stochastic process of daily runoff is transformed into a marked point process (Figure 2-2), defined by the two random variables $n_0(\Delta t)$ (number of peaks in an arbitrary time interval of fixed length Δt) and x (magnitude of the event of discharge or rainfall). We will show that all the information necessary for an analysis of the flood and rainfall extremes is contained in these two variables.
Consider the probability distribution of the magnitude x of the peaks, $F_X(x)$: this distribution can be related to the frequency of occurrence of the peak values. We define first the reference frequency of the flood or rain events, as the expected value of the number of peaks occurring in one year:

(1) $$\Lambda_0 = \mathrm{E}(n_0(\Delta t = 1 \text{ year}))$$

where E( ) indicates expectation. It is then possible to retain only those peaks of the marked point process that exceed a threshold s (see Figure 2-2), and consider their average number in one year, defined as $\Lambda(s)$. Using the same definition given in Equation (1), $\Lambda(s)$ can be denoted as

(2) $$\Lambda(s) = \mathrm{E}(n_s(\Delta t = 1 \text{ year}))$$

where $n_s(\Delta t)$ is the number of marks in Δt with magnitude $x \geq s$. Notice that the $\Lambda(s)$ value in Figure 2-2 results lower for the method M2 than for M1, for any threshold s. The difference is particularly relevant when low thresholds are considered, because the flow component eliminated with the method M2 can represent a significant portion of the measured runoff (see Figure 2-1, peaks between days 30 and 50).
There exist a simple explicit relationship between the frequencies $\Lambda(s)$ and the correspondent CDF's of the magnitudes x: in fact, for time-discrete processes, as in our case, the probability $1-F_X(s)$ that a given event is greater than s is given, by definition, by the number of events exceeding the threshold s divided by the total number of events:

(3) $$\frac{\Lambda(s)}{\Lambda_0} = 1 - F_X(s) = \frac{1}{T(s)}$$

where *T(s)* is the return period of *s* in years. This relationship is fairly general under some basic requirements that are the same as those implying the uniqueness of the CDF: 1) independent mark values; 2) stationary marked point process; 3) identically distributed marks.
The necessity of the third condition is understandable, and requires due care, while the sense of the first two deserve specific attention. The independency issue is crucial, since some confusion may derive from the bivariate nature (i.e. marks and occurrence) of the marked point process defined with the POT procedure.
Consider first the counting process by itself, without consideration of the presence of the marks. A possible representation of the process is as a sequence of experiments (e.g., one for each day when a daily time scale is considered) classified as successes when an event occurs and as failures when it does not. Independency in the occurrences requires that the experiments are independent, i.e. that the probability of success of the *n*-th experiment is not affected from the successes or failures of the previous (*n*-1) experiments. This is called *memoryless property*. However, Equation (3) is valid also for processes that do not possess the memoryless property, because it strictly derives from the definition of the CDF of a time-discrete process. It follows that, even though only few distributions of the number of occurrences $n_0(\Delta t)$ (or $n_s(\Delta t)$) possess the memoryless property (for example the Poisson and the binomial distributions), Equation (3) is valid for any distribution of $n_0(\Delta t)$. For example, in a point process where a success and a failure regularly alternate the occurrences are far from independent, but, provided that the magnitude of the events are independent, Equation (3) is easily seen to be valid. In short, the hyphothesis that the occurrences are Poisson distributed is not necessary for the validity of Equation (3), but it can still be very important for relating the POT distribution, $F_X(x)$, to the corresponding annual maxima distribution, $G_{AM}(x)$. In fact, in case of Poisson distributed occurrences the following relationship holds (e.g., Todorovic, 1978)

(4) $$G_{AM}(x) = e^{-\Lambda_0(1-F_X(x))}$$

From Equation (4) derives that an exponential distribution of the marks produces Gumbel-distributed annual maxima, while the GEV distribution is obtained with Pareto-distributed magnitudes of the events (e.g. Madsen et al., 1997).
Even more important for validity of Equation (3) is the independence among the magnitudes of the events: we clarified in Section 2.1 possible methods to minimize the chance of selecting correlated peaks.
Given independency, the stationarity requirement is not fundamental to the definition of (3), since an analogous equation accounting for the time dependency can be easily written as

(5) $$\frac{\lambda\ (s,t)}{\lambda_0(t)} = 1 - F_X(s,t) \cdot$$

where $\lambda_0(t)$ represents the rate of occurrences of the events, i.e. the mean number of events in the time interval [t,t+Δt], with $\Delta t \to 0$, while $\lambda\ (s,t)$ is the correspondent rate for the events with magnitude exceeding *s.* The relationship between the rates λ and the frequencies Λ is

(6) $$\Lambda(s,t) = \int_t^{t+\Delta t} \lambda\ (s,u)du$$

When the stationarity condition holds, one has $\Lambda(s)=\lambda(s)\Delta t$ and $\Lambda_0=\lambda_0\ \Delta t$, and Δt cancels out from the ratio on the left hand side of Equation (5), giving Equation (3). At a first instance we have assumed stationarity, using Equation (3) at a yearly time scale even in presence of different climatic regimes for the italian rivers (see e.g. Claps and Villani, 2001). Otherwise, one may divide the dataset in *m* parts representing the different climatic regimes and apply Equation (5) to each subset of data: In this second case one ends up with *m* different CDF's of the magnitudes of the peaks. Typically *m*=4 is sufficient even for climates with strong seasonalities: in an application to daily rainfall in Southern Italy, Sirangelo and Iiritano (1999) show that *m*=2 could be a good compromise between detailed time series description and practical calibration. However, when the method M2 is used for peaks selection, a strong seasonal component of the discharges, due for example to snow melting baseflow, is eliminated.

3. ANALYSIS OF FLOOD PRODUCTION MECHANISMS

Peak selection procedures detailed in Section 2.1 are applied in the following to the time series of rainfall *p* and runoff *q* (expressed in *mm*). The average number of rain events in one year defines the value of Λ_{0p} while Λ_{0q} defines the average number of flood events. It is then possible to find the number of rain and flood events with magnitude exceeding *s*, obtaining two different relationships between *s* and the frequencies of occurrence, $\Lambda_p(s)$ and $\Lambda_q(s)$. These relationships, along with the Λ_{0p} and Λ_{0q} values, completely define the empirical CDF's of rainfall and discharge peaks by means of Equation (3). An example of the possible shapes of the above relationships is given in Figure 3-1, where a logarithmic scale is used for the vertical axis to allow a greater detail in the low frequencies zone, more representative for the extreme events, and also to easily recognise if the probability distributions of the magnitudes are exponential, in which case they would plot as straight lines. The two frequency-threshold relations for rainfall and discharge are then compared in order to analyse if there are changes in the flood production mechanisms that can be related to the frequency of occurrence of the events. First, a value λ for the frequency of the events is set; then the corresponding thresholds, s_q and s_p are found for rainfall and runoff (see Figure 3-1).

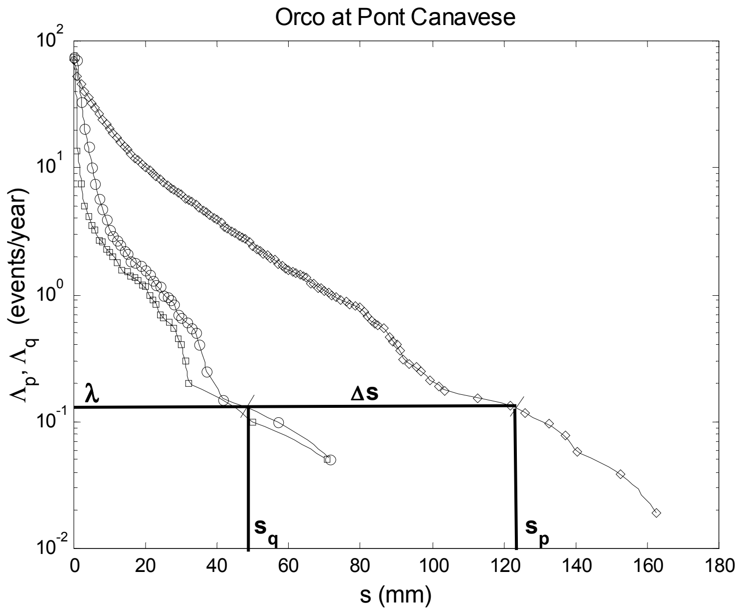

Figure 3-1. Graphical representation of the procedure adopted for obtaining the $\Delta s(\lambda)$ relationship. Diamonds represent the frequency-threshold relation for the rainfall events; circles and squares are related to runoff events obtained with methods M1 and M2 for peaks selection, respectively.

The difference between s_p and s_q represents the rainfall loss for that given frequency, namely

(7) $$\Delta s(\lambda) = s_p(\lambda) - s_q(\lambda)$$

The meaning of $\Delta s(\lambda)$ can be more clear if one considers a simple mechanism of runoff production. Introducing Φ as an abstraction term $(P - Q)$, it can be demonstrated that Φ equals Δs for each λ, provided the rainfall-runoff transformation is deterministic. For this reason, Δs is assumed as a meaningful control variable in the comparison of rainfall and runoff frequency distributions. In addition, consideration of Δs makes it possible the above comparison even when the time series of the two variables (*P* and *Q*) are measured in non-overlapping years, which often happens when short time series of daily rainfall and runoff are available. This is actually the case of the application presented in the following section. Another important advantage of this procedure is that it works at the coarse level

of average values (see Equation (2)), thus eliminating the problem of correctly matching the single rainfall and discharge peaks working at the scale of the event.
Working at equal frequency values λ hampers the usage of an analogous of this procedure with annual maximum data, because in that case the actual frequencies of occurrences are found only in an indirect manner using relations valid in special cases (e.g., Equation (3) and (4) when the occurrences are Poisson distributed).

4. CASE STUDY

35 time series of daily data of runoff and areal rainfall are analysed to evaluate the possibility of highlighting different flood production mechanisms. Drainage basins are mostly located in the North-West of Italy, in the Piemonte region. The Alps influence the majority of basins in the considered group, but some basins are located in a different (Apennine) context. Although not really covering the entire climatic and geologic variety available in Italy, this datasets represent a significant starting point with regard to the mechanisms of flood formation they show.
In some cases, available datasets are not sufficiently long to rely on the resulting information. We decided that reliable data should present a $\Lambda(s)$ relationship comparable in shape to that obtained from the analysis of the annual maxima of rainfall and runoff, as resulting from an independent, and tested, dataset (SIVAPI, 1999).
Evaluating the frequency-threshold diagram ($\Lambda(s),s$) and the corresponding loss-frequency plot ($\Delta s, \Lambda(s)$), we found basin responses which can be grouped as follows:

i. the loss Δs increases steadily with decreasing frequencies (Figure 4-1b);
ii. Δs reaches a maximum for rare events and then decreases (or reaches a plateau) (Figure 4-1d);
iii. the relationship between Δs and frequency presents multiple curvatures (Figure 4-1f).

Trying to understand more quantitatively the physical mechanisms underlying these different patterns, one may analyse simple theoretical models of the rainfall-runoff transformation. The simpler possibility is to consider the flood amount (in mm over a day) produced by a constant abstraction Φ on the rainfall depth P, namely $Q = P - \Phi$. In this case the loss Δs coincides with the abstraction Φ, and remains constants with varying frequencies. This model seems to be able to interpret only small portions of the loss-frequencies relations (e.g., the plateau for intermediate frequencies in Figure 4-1f).
A somewhat more general model considers a linear relationship between rainfall and runoff: $Q = \alpha(P - \Phi)$, where α is a routing coefficient, to be taken in the range [0,1], that depends on climate, size and geomorphological characteristics of the basin. In this case the abstraction Φ does not correspond to the actual loss Δs, and the latter increases with decreasing frequencies, because the difference is proportional to P. All patterns showing increasing Δs up to low frequencies are compatible with such simple model, but the variety of curves found require further analyses on the actual role of the coefficient α.
To discriminate among the above patterns one can also compare the mutual shape of the $\Lambda(s)$ curves for the rainfall and runoff processes. The point is to evaluate if and how these curves follow different patterns and to relate them to the cases described above. We will refer to type **a** curves for those showing a linear shape in the semilog $\Lambda(s)$ diagram, to type **b** for those presenting a marked upward curvature in the diagram (high skewness) and to type **c** curves for the case of downward curvature. Examples of **a** and **b** curves are given by the relationships for precipitation in Figure 4-1a and 4-1e, respectively.
The differences in basin response (i to iii above) can therefore be referred to three main categories of rainfall and runoff curves:
i) Almost-linear response, with both rainfall and runoff following the type **a** pattern. A typical example is given in Figures 4-1a and 4-1b for river Rutor, in Nortwestern Italy, having area of 49 km^2 and average elevation of 2616 m a.s.l. The behaviour here can be interpreted with a linear relationship between precipitation and discharge. Many high-elevation basins among those considered seem to belong to this category. Note the marked difference on the low peaks between curves produced by the M1 and M2 procedures, due to the removal of the glacial baseflow in summer. The case of rainfall and runoff both of type **b** is also present but does not necessarily reflects simple linearity in the response.

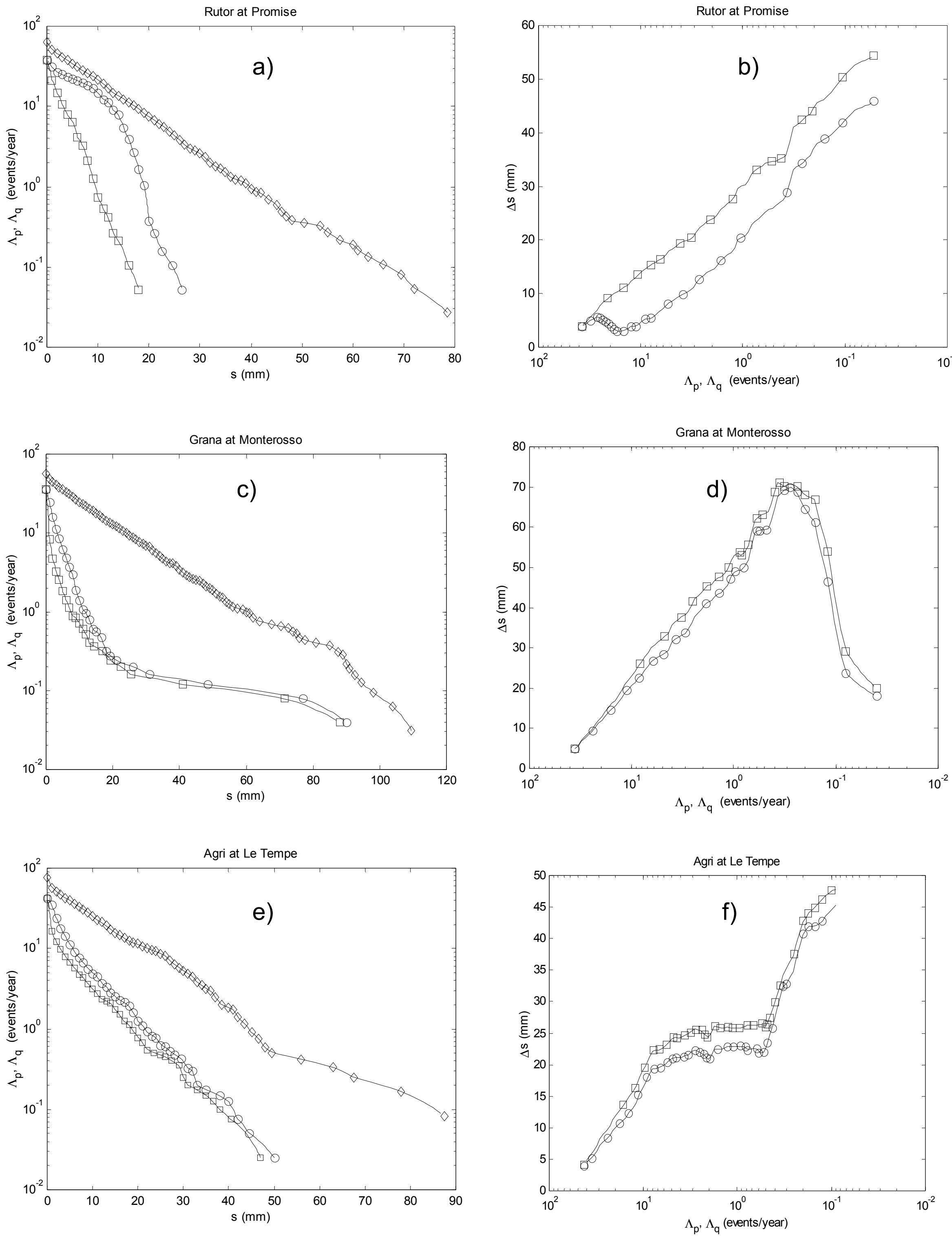

Figure 4-1. Left panels report examples of the frequency-threshold relationships for rainfall (diamonds) and runoff; right panels show relations between the current loss and the discharge frequency (in all cases, circles correspond to method M1 for peaks selection, squares to method M2). Note the reversed scale in the diagrams on the right.

ii) “parabolic” response, with type **a** rainfalls and type **b** runoff. An example is given by river Grana (Figures 4-1c and 4-1d), again in Northwestern Italy (area of 109 km^2 and average elevation of 1540 *m* a.s.l.). A marked change in the floods generation mechanism seems to occur here, with the magnitude of the rare floods that results strongly increased by the reduced absorption of rainfall by the soil. Possible causes for this behaviour are changes in the runoff production mechanism (from an infiltration capacity to a saturation from below mechanism) or differences in the areal coverage of precipitation when “normal” or rare events are considered. The geomorphological and climatic characteristics of the basins belonging to this category are currently under investigation with data from additional basins.

iii) “mixed” nonlinear response, with type **b** rainfalls and type **a** runoff. An example is river Agri (area of 174 km^2 and average elevation of 933 *m* a.s.l.) in Southern Italy (Figures 4-1e and 4-1f). In this case the loss-frequency relation tends to assume one of the shapes described above, but then deviates upward due to the greater intensity of the extreme rain events. The fact that Δs increases again after it has reached stability shows that the capacity of the soil was not exhausted yet notwithstanding the downward curvature of the $\Delta s(\lambda)$ relationship. This requires a very high absorption capacity of the soil, which is found on highly permeable basins (the river Agri basin, for example, is mainly constituted by fractured limestone).

Figure 4-1 deserves further comments. Notice how the shape of the relationships in the right panels tends to be insensitive to the choice of the peak selection procedure (method M1 or M2). This is common to all of the data analysed and is quite important. In fact, we have seen that method M2 minimises the problems deriving from seasonal components and from the possible lack of independence among the marks. The fact that the relations found with method M1 are very similar to those obtained with method M2 demonstrates that, in all the series considered here, seasonality and mutual dependence do not really influence the results. The case of river Rutor, for instance, shows that a marked seasonality in the runoff can produce a deviation only in the lower part of the relationship (Figure 4-1b). The method adopted in this paper seems therefore rather robust.

A second important comment regards the fact that only part of the diagrams in Figures 4-1b, 4-1d and 4-1f are actually meaningful in terms of design floods. In other words, the left part of the diagrams concerns very frequent events, in which the increase of the loss at decreasing frequencies corresponds to the increasing absorption capacity of the soil with not very intense rainfalls. The interesting part of the figures, in terms of annual maxima, are instead on the side of low frequency events (right side). This fact must be carefully considered when using models of transformation that look unsuitable in the side of low frequencies.

5. CONCLUSIONS

The Peak Over Threshold analysis of continuous rainfall and runoff daily data presented in this paper aims to highlight phenomenological mechanisms dominating the flood formation in different physiographic and climatic contexts. An objective procedure was devised for identification of the rainfall and runoff peaks and a simple method for removal of runoff dependence and baseflow effects was proposed. Scope of the analysis is to demonstrate that the relation between the number of flood and rain events exceeding different thresholds can provide meaningful information on the basin transformation. Qualitative information is deduced by drawing diagrams of loss variation versus the frequency of the observed events; quantitative information can be deduced using simple models of the transformation and estimating relevant parameters. The latter step is currently being investigated and is only quickly sketched here.

Markedly different patterns of runoff formation were found on a set of 35 rainfall and runoff time series recorded in Italian rivers, and exploratory justification were provided for their shapes, using also rainfall and runoff frequency curves. These patterns show differences in the mechanisms of basin transformation not easily recognisable from observation of the annual maximum frequency curve. Simple models of constant and proportional abstraction were preliminarily suggested for partial explanation of the observed patterns; derivation of analytical curves from these models under different hypotheses is currently under way.

In view of a thorough specification of the patterns of non-linearity in the rainfall-runoff transformation presented in this paper, climatic issues also require to be integrated in the procedure, being the climatic regime an important source of dissimilarity among basins. Seasonal analyses performed on the same class of data (e.g. Merz et al., 1999; Claps and Villani, 2001) can possibly provide a significant added value to the results shown here.

REFERENCES

Bobée, G. et al. (1993): Towards a systematic approach to comparing distribution used in flood frequency analysis, J. Hydrol., 142, 121-136.

Claps, P., Villani, P. (2001): Using rainfall and runoff peaks over threshold in the analysis of flood generation mechanisms, Proceedings of the 3rd EGS Plinius Conference, held at Baja Sardinia, Italy, October 2001 (in press).

Hashemi, A. M. et al. (2000): Climatic and basin factors affecting the flood frequency curve: PART I -A simple sensitivity analysis based on the continuous simulation approach, Hydrology and Earth System Sciences, vol. 4, no. 3, pp. 463-482.

Iacobellis, V., Fiorentino, M. (2000): Derived distribution of floods based on the concept of partial area coverage with a climatic appeal, Water Resour. Res., 36(2), 469-482.

Lang, M. et al. (1999): Towards operational guidelines for over-threshold modeling, J. Hydrol., 225, 103-117.

Madsen, H. et al. (1997): Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events. 1. At-site modeling, Water Resour. Res., 33(4), 747-757.

Merz, R. et al. (1999): Seasonality of flood processes in Austria. In: “Hydrological Extremes”. Proceedings of the IUGG 99 Symposium, Birmingham. IAHS Publ. No. 255, pp. 273-278.

Robson, A., Reed, D.W. (1999): Flood estimation Handbook. Vol. 3: Statistical procedures for flood frequency estimation, Institute of Hydrology, Wallingford, UK.

Sirangelo, B., Iritano G. (1999): I processi puntuali nello studio delle occorrenze delle precipitazioni intense, L’Acqua, 6, 35-60 (in italian).

SIVAPI (1999): Sistema Informativo per la Valutazione delle piene in Italia, http://www.cs.cnr.it/GNDCI/SivapiNew.htm (date of access, Mar. 2002).

Todorovic, P. (1978): Stochastic models of floods, Water Resour. Res., 14(2), 345-356.