

# Stochastic index model for intermittent regimes: from preliminary analysis to regionalisation

M. Rianna, F. Russo, and F. Napolitano

Dipartimento di Ingegneria Civile, Edile e Ambientale, Sapienza Università di Roma, Via Eudossiana 18, 00184 Roma, Italy

Received: 19 June 2010 – Revised: 24 December 2010 – Accepted: 18 February 2011 – Published: 27 April 2011

**Abstract.** In small and medium-sized basins or in rivers characterized by intermittent discharges, with low or negligible/null observed values for long periods of the year, the correct representation of the discharge regime is important for issues related to water management and to define the amount and quality of water available for irrigation, domestic and recreational uses. In these cases, only one index as a statistical metric is often not enough; it is thus necessary to introduce Flow Duration Curves (FDC).

The aim of this study is therefore to combine a stochastic index flow model capable of reproducing the FDC record period of a river, regardless of the persistence and seasonality of the series, with the theory of total probability in order to calculate how often a river is dry.

The paper draws from preliminary analyses, including a study to estimate the correlation between discharge indicators  $Q_{95}$ ,  $Q_{50}$  and  $Q_1$  (discharges exceeding 95%, 50% or 1% of the time, respectively) and some fundamental characteristics of the basin, as well as to identify homogeneous regions in the target area through the study of several geomorphological features and climatic conditions. The stochastic model was then applied in one of the homogeneous regions that includes intermittent rivers.

Finally, the model was regionalized by means of regression analysis in order to calculate the FDC for ungauged basins; the reliability of this method was tested using jack-knife validation.

water resources management and environmental planning. An important environmental problem related to this issue is wastewater discharge into rivers: in fact, legislation has restricted discharge into rivers characterized by no flow for long periods.

Flow Duration Curves (FDC), which represent the percentage of time during which the discharge of a river is exceeded, can be used as a tool to accurately represent the streamflow frequency regime and can be applied to all these hydrological applications. The FDC can be easily calculated as the complement of the cumulative distribution function (cdf) from a gauged river; this information is also essential for ungauged basins.

FDC characteristics were synthesised by Searcy (1959). Later, Smakhtin (2001) revised the argument on low flows in hydrology, outlining the use of flow duration curves that had been used up to then. FDC have been calculated in ungauged basins in several studies (Quimpo et al., 1983; Mimikou and Kaemaki, 1985; Claps and Fiorentino, 1997; Smakhtin et al., 1997; Ganora et al., 2009); additional studies have provided analysis of the uncertainty of FDC (Yu et al., 2002) and the development of a stochastic model for calculating FDC (LeBoutillier and Waylen, 1993; Cigizoglu and Bayazit, 2000). Vogel and Fennessey (1994) introduced Annually-Based Flow Duration Curves (AFDC), which are useful for making probabilistic considerations of the discharge of dry, wet and average years and for calculating the inter-annual variability associated with AFDC. Castellarin et al. (2004) introduced a similar approach to that of the discharge index to model the relationship between FDC and AFDC in daily discharges. This method can reproduce FDC and also mean, median and variance of AFDC without assumptions based on the seasonal and persistence structure of daily discharges.

These methods do not allow for the calculation of flow duration curves on basins with an intermittent regime, where the above-mentioned environmental problems are critical. These basins are characterized by flash floods and have no

## 1 Introduction

An accurate representation of a river regime is essential to several engineering applications, such as the analysis of hydroelectric feasibility, reservoir and lake sedimentation,



Correspondence to: M. Rianna  
(maura.rianna@uniroma1.it)

base flow, or their base flow is restricted to only the wet periods of the year, while for the rest of the year there is no flow.

The presence of zeros in a discharge time series can be real or can occur when the discharge is beneath a threshold and the instruments cannot take any measurements, as in the case of censored data (Durrans et al., 1999). Several studies have focused on various methods to work with zero data in frequency analysis (e.g. Jennings and Benson, 1969; Kilmartin and Peterson, 1972; Haan, 1977; Wang and Singh, 1995), while others have concentrated more on techniques for working with censored data (e.g. Kroll and Stedinger, 1996; Tate and Freeman, 2000).

The occurrence of zero events can be expressed in probability theory by substituting a non-zero probability mass with a zero value. This creates a discontinuity in the density function from which the hydrological series is obtained, with discontinuity in the zero value. However, this solution can create problems with the assumption of continuity made in frequency analysis. Jennings and Benson (1969) highlighted the potential problems encountered when a continuous distribution is fitted with data including zero values.

The literature presents three methodologies for approaching zero data, summarized in Haan (1977):

- The first is to add a small constant to all observations, such as 1% of the mean magnitude, and fit a continuous distribution such as the Log-Pearson type III distribution onto the data (Subcommittee on Hydrology, 1966). This approach can move the discontinuity represented by the zero state but does not solve the problem created by the discontinuity.
- The second ignores the zero values and considers only the non-zero values and then corrects the results for the entire period recorded. Haan (1977) and Wang and Singh (1995) showed that this method is biased, since it ignores all zero values in the data.
- The third is based on the theorem of total probability (Jennings and Benson, 1969).

Woo and Wu (1989) and Wang and Singh (1995) developed empirical three-parameter models for the frequency analysis of hydrologic data containing zero values starting from the theorem of total probability. On the other hand, Strupczewski et al. (2003) proposed a different method based on the hypothesis that the unit impulse response of a linearized Kinematic Diffusion (KD) model is a probability distribution suitable for frequency analysis of hydrologic samples with zero values.

The principal aim of this paper is to create a model to calculate FDC for daily streamflows that also works in basins with intermittent flow in dry climates. In order to reach this objective, the model combines the stochastic index flow

model (Castellarin et al., 2004) with the theory of total probability. The stochastic index flow model allows for the computation of a river's FDC without regard to the persistence and seasonality of the series and enables the calculation of conditional distribution  $F(y|Y > 0)$ , while the theory of total probability allows an evaluation of the percentage of time the river is dry. A procedure of regionalisation of the model is also applied based on the definition of homogeneous regions in the target area and allows for the definition of equations that permit the transfer of the model to ungauged basins.

This paper is organized as follows:

1. Definition of the modified stochastic index flow model.
2. Presentation of a case study and the regionalisation method.
3. Application of the methodology and discussion of the results.
4. Summary and conclusions.

## 2 The modified stochastic index flow model

### 2.1 Flow duration curves

The FDC provide information about the percentage of time a particular streamflow was exceeded over a given historical period. For a daily flows series, the FDC can be seen as the complement of the cumulative distribution function of daily streamflows based on the complete recording of flows.

A nonparametric approach to construct the FDC can therefore be used (Vogel and Fennessey, 1994):

- By re-assembling the observed streamflows in ascending order.
- By plotting each observation versus its corresponding duration or exceedance probability. Duration is often expressed as a percentage and coincides with an estimate of the exceedance probability,  $e_i$ , of the  $i$ -th observation in the ordered sample.  $e_i$  can be estimated using an empirical distribution such as the Weibull plotting position. Duration  $D_i$  is thus

$$D_i = 100(e_i) = 100 \cdot \left(1 - \frac{i}{n+1}\right),$$

for  $i = 1, 2, 3, \dots, n$  (1)

where  $n$  is the length of the sample.

It is possible to build FDC with different time resolutions of the discharges, but daily FDC offer the most detailed means of examining the duration characteristics of a flow.

An approach to modelling daily streamflows and estimating FDC is based on the stochastic index flow model (Castellarin et al., 2004), which is similar to the flood index approach (Dalrymple, 1960).

### 2.2 Standard stochastic index flow model of FDC

The approach used here assumes that daily streamflow  $X$  can be found by multiplying an index flow equal to annual flow AF by a dimensionless daily streamflow  $X'$ ,

$$X = AF \cdot X' \tag{2}$$

The climatic conditions and annual precipitation given for a basin affect AF. The probability density function,  $f_{X'}$  of standardized flows is correlated with the geomorphologic characteristics of the basin.

Using this formulation, it is possible to calculate FDC for the complete recording period of flows as the complement of the cumulative distribution function (cdf) of  $X$ ,  $F_X$  given by:

$$\begin{aligned} F_X(x) &= \Pr\{X \leq x\} = \int_{x_l}^x f_X(u) du = \Pr\{AF \cdot X' \leq x\} \\ &= \int_{\Omega_{X'}} \int_{af_l}^{x/z} f_{AF,X'}(v, z) dv dz \end{aligned} \tag{3}$$

$\Omega_Y$  = domain of a given random variable  $Y$ ;  $f_X$  = pdf of  $X$ ;  $f_{AF,X'}$  = joint probability distribution of AF and  $X'$ ;  $x_l$  and  $af_l$  = lower bounds of  $\Omega_{X'}$  and  $\Omega_{AF}$ , respectively.

If it is assumed that AF and  $X'$  are independent, then  $f_{AF,X'}$  equals the product of the two marginal distributions, and it is possible to write:

$$F_X(x) = \int_{\Omega_{X'}} f_{X'}(z) \int_{af_l}^{x/z} f_{AF}(v) dv dz = \int_{\Omega_{X'}} f_{X'}(z) F_{AF}(x/z) dz \tag{4}$$

where  $F_{AF}$  = cdf of AF;  $f_{X'}$  = pdf of  $X'$ .

The FDC can be estimated by plotting the variable  $X$  against the duration, equal to  $100(1 - F_X)$  (Castellarin et al., 2004).

### 2.3 Stochastic index flow model in the presence of zero data

The problem of the presence of zero data can be solved using the theorem of total probability, which is used to determine the probability of occurrence of a non-zero event, given that a zero event has already occurred (Jennings and Benson, 1969).

The theorem is given by:

$$\begin{aligned} \Pr(X > x) &= \Pr(X > x | X = 0) \Pr(X = 0) \\ &+ \Pr(X > x | X \neq 0) \Pr(X \neq 0) \end{aligned} \tag{5}$$

Thus, given that  $\Pr(X > x | X = 0)$  is zero, and writing this relationship in the form of cumulative probability distributions, it is possible to obtain  $\Pr(X \leq x)$ :

$$\Pr(X \leq x) = p_{dry} + p_{nz} \Pr(X \leq x | X \neq 0) \tag{6}$$

where  $p_{nz}$  = the percentage of time that the river is flowing (i.e.  $\Pr(X \neq 0)$ ).  $p_{nz}$  can be estimated using the plotting position formulation;  $p_{dry}$  = the percentage of time the river is dry, equal to  $1 - p_{nz}$ .

Therefore the conditional distribution  $\Pr(X \leq x | X \neq 0)$  can be calculated using the stochastic index flow model:

$$\Pr(X \leq x | X \neq 0) = \int_{\Omega_{Xnz'}} f_{Xnz'}(z) \cdot F_{AFnz}(x/z) dz \tag{7}$$

$f_{Xnz'}$  = probability density function of non-zero  $X'$  values;  $F_{AFnz}$  = cumulative distribution function of non-zero AF values.

The calculation of conditional distribution  $F_{AFnz} = \Pr(AF \leq af | AF \neq 0)$  and  $f_{X'nz} = \Pr(X' \leq x' | X' \neq 0)$  with positive values of the series, carried out using a fitting procedure. The empirical frequency distribution conditioned by  $AF > 0$  and  $X' > 0$  can be calculated on non-zero values using a modified Weibull plotting position (Wang and Singh, 1995). In fact, it is possible to consider a situation in which the observed ordered-time series has size  $n(y_1, \dots, y_k, 0, \dots, 0)$  in ascending order of magnitude, where  $y_1, \dots, y_k$ , are all positive, while the other  $n - k$  are zero values. To calculate the Weibull plotting position, it is not possible to use all the  $n$  values and the formulation  $100(1 - m/n + 1)$  for  $m = 1, \dots, n$ , but it is necessary to use the formulation with only the  $k$  positive values:

$$D_i = 100(1 - i/k + 1), \text{ for } i = 1, \dots, k \tag{8}$$

The general formulation of the stochastic index flow model for use with zero values is obtained by incorporating Eq. (7) into Eq. (6):

$$\Pr(X \leq x) = p_{dry} + p_{nz} \cdot \int_{\Omega_{Xnz'}} f_{Xnz'}(z) \cdot F_{AFnz}(x/z) dz. \tag{9}$$

## 3 Case study

The basins of the target area include the Tiber basin as well as its sub-basins, the basins located in northern Lazio and those located north and south of the Tiber River. These areas show considerable lithological variability affecting the large geomorphological structures.

The geology of the area is connected to volcanic activity, since one-third of the region is covered with this kind of geological substrate. The volcanic area has developed in a NW-SE direction, while the northwestern part of the target area involves a metamorphic substrate dating back to the Paleozoic period. The other important geological domain is the Apennine dorsal positioned in the Southeast of the region. This macro system is mainly composed of carbonate sediments dating back to the Mesozoic period and deposited in different sedimentation environments. Most recent sediments in the area are quaternary deposits that make up coastal plains and river valleys.

The geomorphologic characteristics of the region are closely connected to the geological domains. In fact, the big geomorphologic domains coincide with recognized geological structures: the big volcanic districts, the Apennine dorsal, the coastal plains and the remaining Tiber valley. In these big geomorphologic structures, it is possible to identify uniformity and therefore distinctive morphotypes.

The hydrographic structure is controlled by the Tiber River system in the northern part of the region and by the river systems in the South of Tiber. The Tiber River basin covers about 17 200 km<sup>2</sup> and represents the main watercourse of the area.

The Tiber River has an Apennine trend in its initial reach and flows with a torrential regime. Along its right bank, the river collects water from different volcanic districts. From its left bank, it receives water from the Apennine carbonate structures. These contributions stabilize the regime. The Tiber River shows a large difference in hydrographic structure between the basins belonging to the right and left banks. This difference is due to the different ways in which the volcanic systems move, compared to those of carbonate structures characterized by a lower drainage density.

The river basins in the southern part of the study area make up about 4 900 km<sup>2</sup> and, with the Tiber River, supply the area with 80% of the total runoff. Within these river basins, the permeability characteristics and morpho-topographic structures, mainly represented by carbonate platform deposits, determine the highly effective infiltration and consequently slow development of the hydrographic network and low overland flow.

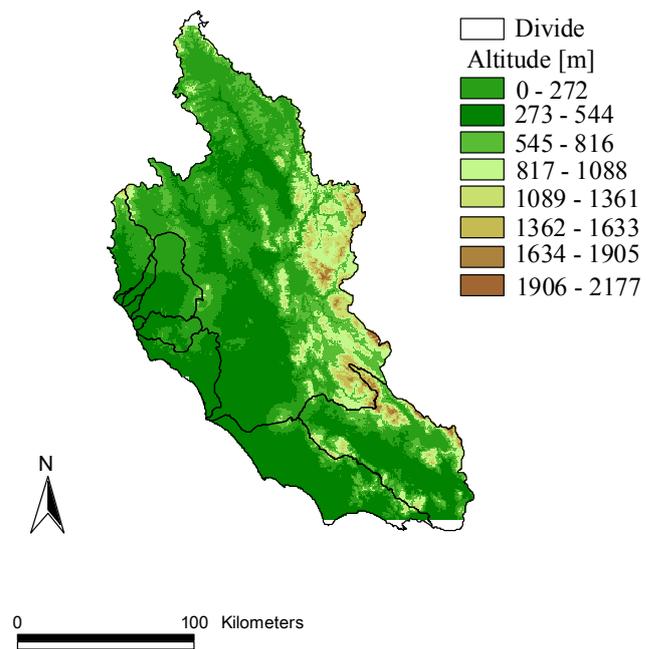
The river basins in northern Lazio have been formed on geological formations with low permeability and a hydrological regime characterized by high overland flow from autumn to winter, when their discharges are 3–4 times higher than those in summer.

The karst system is also particularly developed both in the mountains in the north-east area and in the Apennine dorsal, where there are mostly extended karst shapes of large dimension. Moreover, this area has a highly variable climate due to two major bio-climate regions, temperate and Mediterranean, and the relative transitional regions.

The significant differences in the study area are highlighted in the map of the digital elevation model in Fig. 1, where it is possible to recognize the main basins of the region.

The hydrological data used for the analysis came from the Ufficio Idrografico e Mareografico of Lazio Region. At least 6 years' worth of daily recorded discharges from 26 stations in the study area were used.

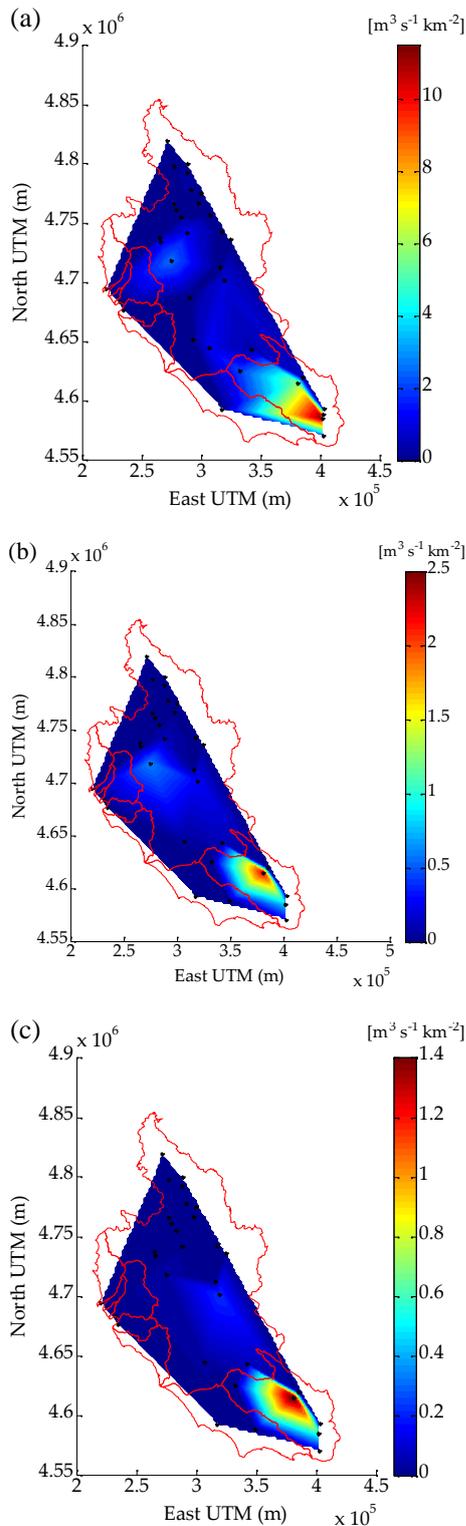
Quantiles  $Q_{95}$ ,  $Q_{50}$  and  $Q_1$ , were estimated. In particular, the minimum discharge  $Q_{95}$  is widely used in Europe and was chosen because of its importance for many applications relating to water management, as in Gustard et al. (1992), Smakhtin (2001) and Laaha and Blöschl (2007).



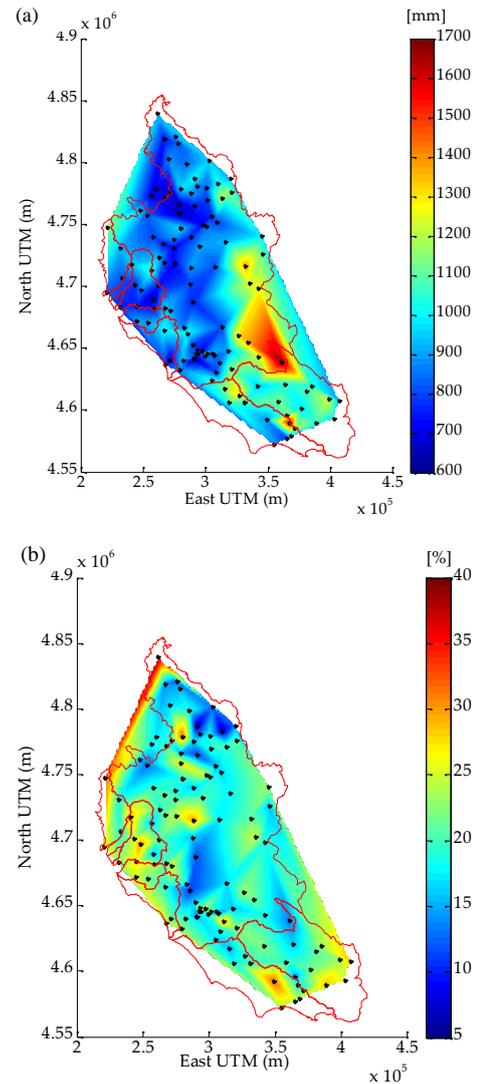
**Fig. 1.** Digital elevation model of the study area. The main basins in the area are divided with black lines.

Each station selected for this study was considered as a separate basin. Thus for basins that do not have upstream stations, the quantiles were calculated directly from the discharge data, by adapting the data of each station to the best possible distribution and then calculating the relative quantiles. The nested basins were, however, divided into sub-basins separated by several measurement stations; the quantiles were calculated as the difference between the quantiles of the upstream and downstream stations. This last estimation is more robust than the quantile calculated from the difference between the hydrographs, but it requires the mean maximum and minimum flow rates to be synchronized in different stations. Furthermore, this method reduces the spatial dependence of discharge data (Laaha and Blöschl, 2007). It is necessary to bear in mind that, in the absence of isochrones, error may also be high. All quantiles  $Q_{95}$ ,  $Q_{50}$  and  $Q_1$  have also been standardized with respect to the area of the sub-basin to obtain the specific quantiles  $q_{95}$ ,  $q_{50}$  and  $q_1$  respectively [ $\text{m}^3 \text{s}^{-1} \text{km}^{-2}$ ]. Figure 2 shows the geographical representation of quantiles  $q_1$ ,  $q_{50}$  and  $q_{95}$ , where it is evident that the largest values of specific quantiles are in the South-east of the area that coincides with the Apennine dorsal.

In addition, the climatic characteristics were calculated by means of rainfall data from 118 rain gauges over the period 1985–2009. The average annual rainfall over the basin was calculated using the Thiessen method. Finally, the coefficient of variation of annual precipitation was calculated and then used to represent the temporal variability of rainfall in the area. Figure 3 represents the mean annual precipitation on each sub-basin and the coefficient of variation of annual



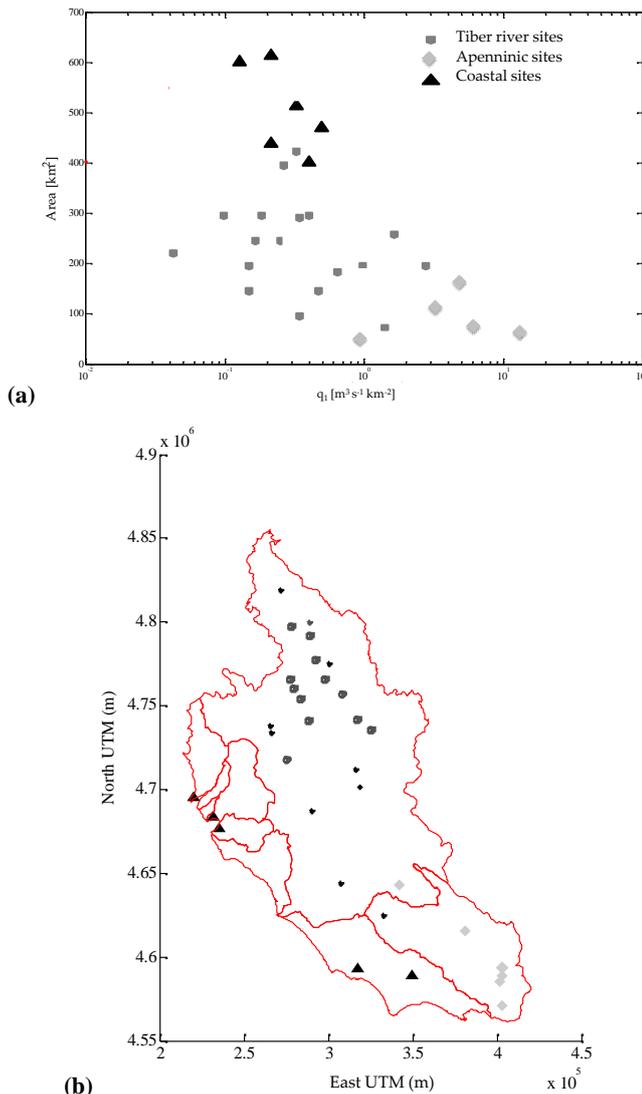
**Fig. 2.** Geographical representation of specific quantiles  $q_1$  (a),  $q_{50}$  (b) and  $q_{95}$  (c). The biggest values of quantiles are located in the Southeast (Apennine area), and the dimension of the region of interest decreases from higher ( $q_1$ ) to lower quantiles ( $q_{95}$ ). The black points in the figures represent the gauge stations in the study region.



**Fig. 3.** Geographical representation of the mean annual precipitation MAP (a) and of the coefficient of variation CV of annual precipitation (b). The black points represent the locations of rain gauges in the region.

precipitation. Again, in this case the orography influences the amount of rainfall over the basin, while the coefficient of variation is lower in the Apennines and higher in the coastal areas. The layout of the digital terrain model, precipitation, variation coefficient and quantiles of the discharge (in particular the maximum) highlight how strongly the orography influences the hydrological behaviour of the region.

Figure 4 shows the basin area against the specific quantile of discharge  $q_1$ . This scatter plot was used as a preliminary qualitative analysis to individuate the presence of structures in the data. Using the scatter plot, it was possible to identify the presence of three data sets: Coastal, Apennines and Tiber basin stations were identified. The same type of behaviour is present in the scatter plot of the quantiles  $q_{50}$  and  $q_{95}$ .



**Fig. 4.** The scatter plot of basin area against specific quantile of discharge  $q_1$  used as qualitative analysis for preliminary identification of structures in the dataset (a) (semi-logarithmic representation). It is possible to recognize three different regions. In the figure below (b), points of the scatter plot are also identified geographically. The geographical representation points represent the flow level stations.

These groups are formed according to the climatic conditions of the area. In particular, it is possible to note how the stations characterized by greater discharges are those in the Apennine region also characterized by higher precipitation depths.

## 4 Application and results

### 4.1 Selection of geomorphological variables and correlation analysis

Correlation analysis allows for the evaluation of variables that influence quantiles in the target area.

**Table 1.** Predicting variables and annotation.

Variable	Description	Unit
A	Area	km <sup>2</sup>
H <sub>mean</sub>	Mean Altitude	m
MAP	Mean annual precipitation	mm
CV	Coefficient of Variation of precipitation	%
F <sub>A</sub>	Calcareous substrate	%
F <sub>D</sub>	Impervious substrate	%
G <sub>A</sub>	Agricultural areas	%
G <sub>B</sub>	Forested areas	%

The choice of characteristics for the basins that were considered for this study was guided by the analysis of the interactions among the flow regime, climate and physical characteristics.

The geomorphological characteristics were calculated for each basin, using GIS data to obtain the area and maximum height of the basin. The Corine Land Cover (CLC) map, taken from the CORINE program in Lazio and Umbria, was used to evaluate soil use. The hydrogeological map of the area was used to define the percentages of different lithological structures for each basin.

Table 1 shows the variables as well as the symbols used. Table 2 shows the minimum, average and maximum values of some geomorphoclimatic indexes. The characteristics considered are sub-basin area A (km<sup>2</sup>); the maximum, average and minimum elevation of the basin (H<sub>max</sub>, H<sub>mean</sub>, H<sub>min</sub>) in meters above sea level (m a.s.l.); the value  $\Delta H = H_{\text{mean}} - H_{\text{min}}$  in meters; the percentage of impervious substrate in the basins (F<sub>D</sub>); and Mean Annual Precipitation MAP (mm) calculated for each basin. The values in Table 2 demonstrate the high heterogeneity and complexity of the study region.

Table 3 shows the correlation matrix estimated for all the dependent variables with data from the study region. Table 4 shows p-values at the 0.05% level when testing the hypothesis of no correlation; small p-values indicate significant correlation. Table 3 shows that the variables most highly correlated with specific quantiles are area, elevation of the basin and lithological characteristics. The correlation of specific quantiles with the area is particularly important. It can be noted that the coefficient diminishes from minimum to maximum discharges. One hypothesis is that the maximum discharges are influenced not only by the basin area but also by the average rainfall, which causes high flow.

The average annual rainfall influences the maximum discharges, which increase as expected when rainfall increases. The coefficient of variation does not appear to affect the discharges greatly. The lower discharges are strongly linked to the lithological characteristics of the soil (Nikic and Radonja, 2009).

In particular, the percentage of carbonate and impervious substrate and thus the permeability characteristics of the basins are strongly linked to minimum discharges. This

**Table 2.** Minimum, average and maximum values of the geomorphological and climatic indexes for the basins in the study region.

	A (km <sup>2</sup> )	H <sub>min</sub> (m)	H <sub>max</sub> (m)	H <sub>mean</sub> (m)	ΔH (m)	F <sub>D</sub> (%)	MAP (mm)
Minimum	31.08	2.00	389.00	32.30	31.30	0.00	650.00
Average	400.05	106.40	1414.60	479.70	396.50	6.53	1066.15
Maximum	981.23	368.00	2200.00	2031.00	2012.00	65.26	1350.00

**Table 3.** Correlation matrix between discharge quantiles and geomorphological characteristics.

	q1	q50	q95	A	H <sub>mean</sub>	MAP	CV	F <sub>A</sub>	F <sub>D</sub>	G <sub>A</sub>	G <sub>B</sub>
q1											
q50	0.75										
q95	-0.34	-0.23									
A	-0.45	-0.48	-0.49								
H <sub>mean</sub>	0.42	0.33	0.31	-0.08							
MAP	0.34	0.01	0.05	0.00	0.13						
CV	-0.09	-0.13	0.02	0.19	0.33	0.12					
F <sub>A</sub>	0.58	0.08	0.44	0.11	-0.17	-0.29	0.04				
F <sub>D</sub>	0.20	0.13	-0.44	0.32	-0.11	-0.62	-0.19	0.42			
G <sub>A</sub>	0.21	0.54	0.16	0.22	0.02	0.41	0.52	-0.04	0.22		
G <sub>B</sub>	-0.41	-0.26	0.12	-0.32	-0.29	0.12	-0.23	-0.42	0.16	0.02	

**Table 4.** p-value matrix to test the hypothesis of no correlation (critical value equal to 0.05%).

	q1	q50	q95	A	H <sub>mean</sub>	MAP	CV	F <sub>A</sub>	F <sub>D</sub>	G <sub>A</sub>	G <sub>B</sub>
q1											
q50	0.02										
q95	0.25	0.43									
A	0.12	0.08	0.09								
H <sub>mean</sub>	0.13	0.25	0.26	0.79							
MAP	0.19	0.98	0.94	0.99	0.66						
CV	0.79	0.66	0.98	0.51	0.25	0.67					
F <sub>A</sub>	0.05	0.79	0.11	0.72	0.54	0.32	0.95				
F <sub>D</sub>	0.46	0.66	0.11	0.26	0.72	0.01	0.51	0.13			
G <sub>A</sub>	0.45	0.05	0.56	0.45	0.98	0.13	0.06	0.95	0.43		
G <sub>B</sub>	0.13	0.37	0.67	0.25	0.32	0.67	0.39	0.12	0.56	0.98	

is probably because higher minimum discharges occur in basins with groundwater flow even in summer. This would also explain the negative correlation with the percentage of impermeable substratum F<sub>D</sub>. Altitude, on the other hand, mainly affects maximum discharges. In particular, it was noted that the stations with the highest discharges are those in the Apennines, which are characterised by higher precipitation depths. This can be explained by the phenomenon of orographic rainfall.

Furthermore, Tables 3 and 4 show large p-values and thus low correlations, probably caused by the wide variability of the study area.

## 4.2 Cluster analysis through homogeneity test

Hierarchical cluster analysis is used to identify relatively homogeneous groups of variables based on selected characteristics, using Ward's algorithm (Ward, 1963), the most commonly used agglomerative clustering technique in the regionalisation context. This technique produces spherical clusters that are all approximately the same size. The aim is to link various objects into clusters using a measure of similarity or distance. This approach begins with  $n$  groups, each of which contains one case. Two of the cases are then combined to obtain a single cluster. At the next stage, either a third case is added to the cluster or two other cases are merged into a new cluster. Ward's method unites the two groups in order to minimise the error sum of squares.

It is also necessary to test whether the data observed at different sites in a homogeneous region arise from a common regional distribution. If the test fails, the association with the region is reconsidered and the procedure is repeated until the region can be considered homogeneous. The two homogeneity tests used were developed by Hosking and Wallis (1997) and estimate the degree of heterogeneity of a group of sites in order to evaluate whether they can be considered homogeneous. The tests are based on the L-moment ratios ( $L_{CV}$ , L-skewness and L-kurtosis) defined by Hosking and Wallis (1997).

The first heterogeneity measure is calculated as:

$$H_1 = \frac{(V - \mu_V)}{\sigma_V} \quad (10)$$

The  $H_1$  measure is based on the sample variance of L-moment ratio  $L_{CV}$ , which Hosking and Wallis (1997) define as the most significant parameter to individuate homogeneous regions and here is identified as  $V$ .

The parameter  $V$  in the  $H_1$  formulation can be calculated as:

$$V = \frac{\sum n_i (L_{cv}^i - \bar{L}_{cv})^2}{\sum n_i} \quad (11)$$

where  $n_i$  is the number of observations in station  $i$ .  $L_{cv}^i$  and  $\bar{L}_{cv}$  are the  $L_{CV}$  of station  $i$  and the mean regional  $L_{CV}$ .

The mean  $\mu_V$  and standard deviation  $\sigma_V$  of the chosen dispersion measure are estimated using this procedure: the mean regional L-moment ratios are used to evaluate the parameters of a kappa distribution. This allows for the calculation of the repeated simulation of a homogeneous region in which the recorded lengths of its sites are the same as those of the observed data. In this case 500 homogeneous regions were generated. Mean  $\mu_V$  and standard deviation  $\sigma_V$  are then obtained from these simulations.

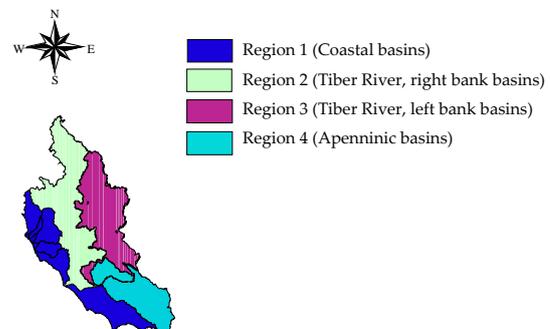
The region can be assumed to be homogeneous if the  $H_1$  is sufficiently small. Hosking and Wallis (1997) suggest that the region may be assumed to be “acceptably homogeneous” if  $H_1 < 1$ , “possibly homogenous” if  $1 < H_1 < 2$  and “definitively heterogeneous” if  $H_1 > 2$ .

The  $H_1$  only measures heterogeneity in the dispersion of the samples, since it is based solely on the differences between the sample  $L_{CV}$  in the region. Hosking and Wallis (1988) also give an alternative heterogeneity measurement, which we call  $H_2$ . It is obtained using the same procedure as that of  $H_1$  measurement but is based on  $L_{CV}$  and L-skewness at the same time.  $H_2$  has similar acceptability limits as the  $H_1$  statistic. Hosking and Wallis (1997) judge  $H_2$  to be inferior to  $H_1$ , stating that it rarely yields values larger than 2 even for highly heterogeneous regions.

Cluster analysis was then applied, using explanatory variables with the higher correlation coefficient. First of all, the basins’ area, altitudes and geographical coordinates were used to cluster sites. In this way three regions were obtained,

**Table 5.** Results of the Hosking and Wallis (1997) homogeneity tests for  $Q_1$ , mean values  $Q_{50}$  and annual minima  $Q_{95}$  for the three regions initially identified.

	$H_1$	$H_2$
Tiber $Q_1$	4.008	0.938
Tiber $Q_{50}$	12.538	1.103
Tiber $Q_{95}$	1.798	1.815
Coastal $Q_1$	0.043	0.536
Coastal $Q_{50}$	3.127	0.407
Coastal $Q_{95}$	2.060	0.649
Appenninic $Q_1$	0.325	0.236
Appenninic $Q_{50}$	3.003	0.395
Appenninic $Q_{95}$	3.313	1.178



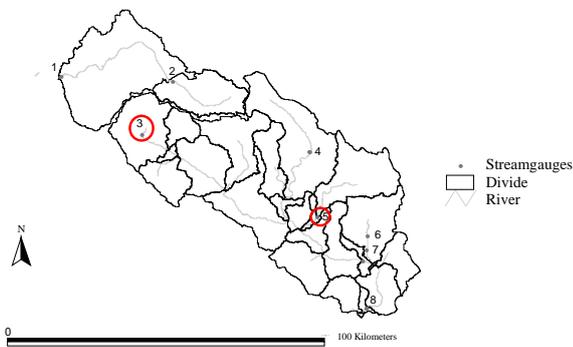
**Fig. 5.** Homogeneous regions recognized in the study area. Four regions were determined according to the geomorphological characteristics of the area.

which coincided with the Apenninic, coastal and Tiber River stations as already identified from the scatter plot. The Hosking and Wallis (1997) homogeneity tests were applied for these regions to the  $Q_1$  values, the annual mean values and  $Q_{95}$  values. The results of these tests are depicted in Table 5, where it can be seen that the more constraining  $H_1$  statistic values are often bigger than the threshold value that identifies a heterogeneous region. A very high heterogeneity was detected in the Tiber River region in particular. To solve this problem, the percentage of substrate (volcanic or carbonatic) was added to the other variables used to cluster sites and a different configuration of the regions was hypothesized. After this procedure the Tiber River basins were divided into two regions: the left bank and the right bank of the river. The other two regions coincide with those initially identified.

Figure 5 shows the four regions identified by cluster analysis, while Table 6 shows the results of the two tests for the Tiber River basins after the division. The values for the  $H_1$  statistic for the two Tiber regions are lower than in the first configuration; the  $H_2$  statistic is less than 2 in all cases. The obtained results are influenced by intersite correlation due to the nested structure of the region, although Hosking and

**Table 6.** Results of the Hosking and Wallis (1997) homogeneity tests for annual maxima  $Q_1$ , mean values  $Q_{50}$  and annual minima  $Q_{95}$  for the two Tiber River regions.

	$H_1$	$H_2$
Tiber Carbonatic $Q_1$	1.246	0.582
Tiber Carbonatic $Q_{50}$	2.632	1.239
Tiber Carbonatic $Q_{95}$	1.842	0.812
Tiber Volcanic $Q_1$	0.124	0.423
Tiber Volcanic $Q_{50}$	0.764	0.512
Tiber Volcanic $Q_{95}$	1.314	0.981



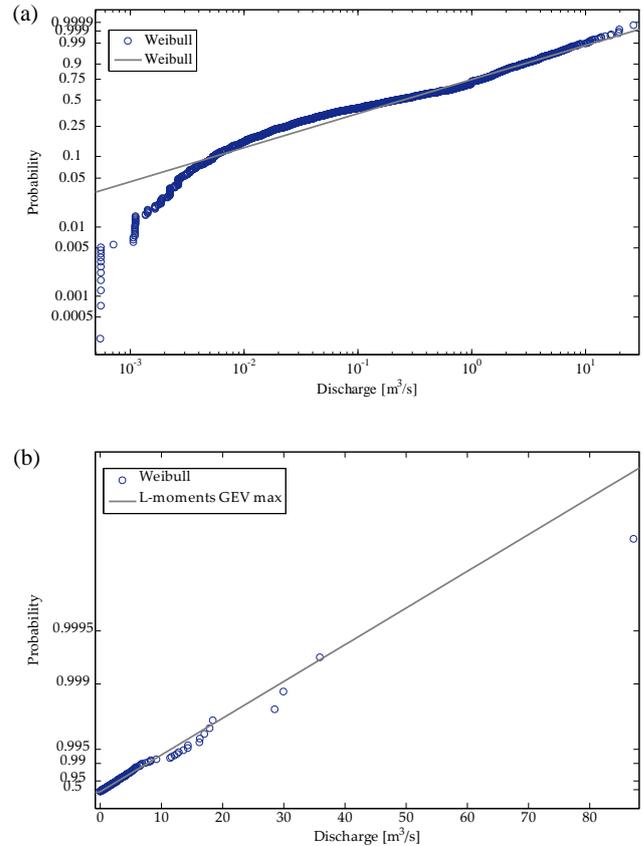
**Fig. 6.** Region 4 (Apennine basins), with gauged sites used to regionalize the stochastic index flow model. The red circles highlight the position of the stations used for the testing of the modified stochastic index flow model.

Wallis (1988) showed that regional heterogeneity affects the accuracy of regional flood frequency quantiles more significantly than does intersite correlation. Hence, because of the great heterogeneity of the region, the tests seem to have been passed.

**4.3 Results for the modified stochastic index flow model**

A more thorough description is shown here for the procedure of the modified stochastic index flow model. Only two stations (nos. 3 and 5) with an intermittent regime were available in the study region, both of which were located in the Region 4 (Apennine basins). Figure 6 shows the Apennine area and the two sites considered in the analysis. Seven years of data, from 2003 to 2010, are available for sites nos. 3 and 5. Initially AF and  $X'$  were calculated for each time series, and several distributions were fitted to these series to calculate the FDC using the stochastic index flow model.

First the FDC was calculated using the stochastic index flow model, without using the theory of total probability. Then the new model was used to calculate FDC when zero data were present. The probability of non-zero flow  $p_{nz}$  and the complement  $p_{dry}$  were calculated for both sites and then zero data were separated from the time series.  $p_{nz}$  was 95%



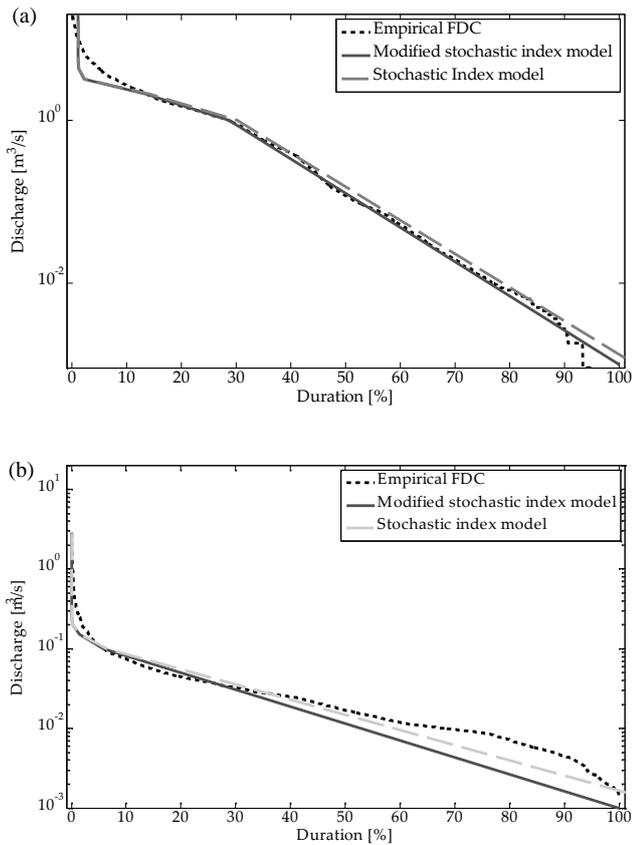
**Fig. 7.** Fitted Conditional Frequency Distribution  $F_{X'_{nz}} = \Pr(X' \leq x' | X' \neq 0)$  for sites n. 5 (a) (logarithmic representation) and n. 3 (b). The best distributions are the Weibull distribution for site n. 5 (a) and GEV max with parameters obtained with the L-moment method for site n. 3 (b).

and 97% for sites nos. 3 and 5, respectively. These values were calculated using the standard plotting position formulation.

Subsequently, a fitting procedure was applied to the non-zero data. Empirical distribution for AF and  $X'$  was calculated using the formulation that considers only non-zero data. Normal distribution is the best distribution for AF non-zero data. Figure 7 represents the fitting of distributions for  $X'$  non-zero data. The distributions that fit the values best were the Weibull distribution for  $X'$  non-zero data from station no. 5 and GEV Max with parameters obtained using L-moments for  $X'$  data without zero from station no. 3.

Computation of conditional distribution  $F_{AF_{nz}} = \Pr(AF \leq af | AF \neq 0)$  and  $f_{X'_{nz}} = \Pr(X' \leq x' | X' \neq 0)$  with positive values of the series was then performed.

It is then possible to calculate the probability of all the data using the modified stochastic index flow model equation (Eq. 9).



**Fig. 8.** Semi-logarithmic representation of fitted FDC ( $\Pr(X > x)$ ) for sites n. 5 (a) and n. 3 (b). The broken black line is the empirical FDC, obtained through the Weibull plotting position, the solid grey line is calculated using the modified stochastic index flow model and the broken bright grey line is FDC calculated using the standard stochastic index flow model.

Figure 8 represents the FDC for the two stations. The results are represented on a semi-logarithmic scale.

In order to evaluate the accuracy of the modified model, the Root Mean Square Error (RMSE) and Nash-Sutcliffe (NS) efficiency coefficient were calculated.

The formulation of RMSE is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Q_o^t - Q_m^t)^2}{n}} \quad (12)$$

where  $Q_o^t$  is the observed discharge at the time  $t$ , and  $Q_m^t$  is modelled discharge at time  $t$ .

The formulation of the Nash-Sutcliffe efficiency coefficient is:

$$NS = 1 - \frac{\sum_{t=1}^T (Q_o^t - Q_m^t)^2}{\sum_{t=1}^T (Q_o^t - \bar{Q}_o)^2} \quad (13)$$

where  $\bar{Q}_o$  is the mean value of the observed discharges.

**Table 7.** RMSE and Nash-Sutcliffe efficiency calculated for the two stations with intermittent regime and for the standard and modified formulation of the stochastic index flow model. The Nash-Sutcliffe efficiency is calculated on logarithms of data.

Site	RMSE		Nash-Sutcliffe efficiency	
	SI [m³ s⁻¹]	modified SI [m³ s⁻¹]	SI	modified SI
5	6.10	5.20	0.94	0.95
3	5.70	7.00	0.96	0.94

Nash-Sutcliffe efficiencies can range from  $-\infty$  to 1. An efficiency value of 1 ( $NS = 1$ ) represents a perfect match of the modelled discharges to the observed data. An efficiency value of 0 ( $NS = 0$ ) indicates that the model’s predictions are as accurate as the mean of the observed data, whereas an efficiency less than zero ( $NS < 0$ ) occurs when the observed mean is a better predictor than the model.

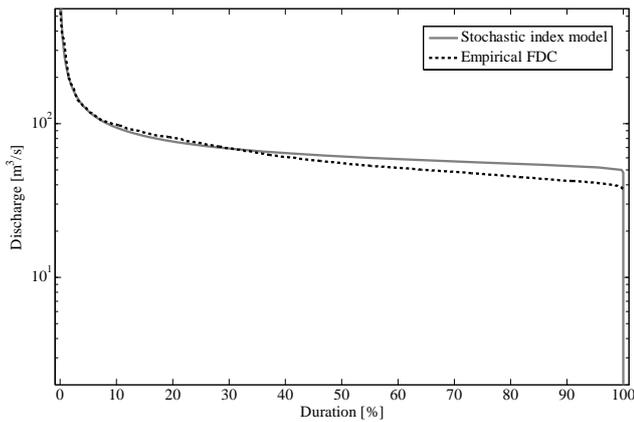
Results for RMSE and NS efficiency are shown in Table 7 for the stations with an intermittent regime. These efficiency indexes are calculated for the FDC using both the stochastic index flow model in the standard formulation and the modified model. The results from the modified stochastic index flow model are better than those from the standard formulation for the station no. 5, while they are worse for station no. 3. The modest difference between the two curves (calculated using the standard and the modified formulations) depends on the small percentage of zero values in the dataset. Moreover, the performance of the modified stochastic index flow model for station no. 3 is probably affected by the particular shape of the empirical curve, due to over-abstraction of groundwater.

#### 4.4 Regionalisation analysis

##### 4.4.1 Choice of the best parent distribution for AF and $X'$

To develop a regionalisation model, all the sites must have the same parent distribution; it is then necessary to choose distributions that closely fit AF and  $X'$  data for all stations. It is also important to choose the distributions with the fewest parameters in order to have a parsimonious model.

The analysis was carried out in Region 4, which corresponds to the Apennine region and comprises eight stations (Fig. 6). The regionalisation approach was tested in this region involving basins with intermittent flow. Six out of the eight sub-basins are considered to have a permanent regime. For this reason, the stochastic index flow model was applied to these stations with the classical implementation. This means that for these stations the value of  $p_{nz}$  is 1 because there is always flow in the river. Figure 9 shows the results of implementing the stochastic index flow model for site no. 8.



**Fig. 9.** Results of application of the stochastic index flow model for site n. 8 (semi-logarithmic representation). The broken black line represents the empirical FDC; the grey line represents the FDC obtained with the stochastic index flow model.

For AF values of the different sites, several distributions have a good fit. The best distribution for the data was the normal one. This result was expected because of the low skewness of the annual flow and the central limit theorem. Moreover, this distribution has only two parameters, the mean  $\mu$  and standard deviation  $\sigma$ .

To evaluate the best distribution for  $X'$  data, a method based on the L-moment ratio diagram that represents L-skewness versus L-kurtosis was used. Figure 10 demonstrates that the GEV distribution gives the best fit for almost all the stations. Observed parameters from the sites of the region are provided in Table 8.

A regression model was used to transfer model parameters to ungauged basins (see Sect. 4.4.2). Parameters of normal distribution as well as scale parameter  $\lambda$  and shape parameter  $\kappa$  of the GEV distribution are estimated by equations dependent on geomorphological characteristics. It is possible to evaluate location parameter  $\xi$  of the GEV distribution using L-moment formulation (Hosking and Wallis, 1997).  $\lambda_1$ , the mean value of the standardized discharge sample  $X'$ , is equal to unity. Location parameter  $\xi$  can thus be evaluated after the estimation of  $\kappa$  and  $\lambda$  parameters:

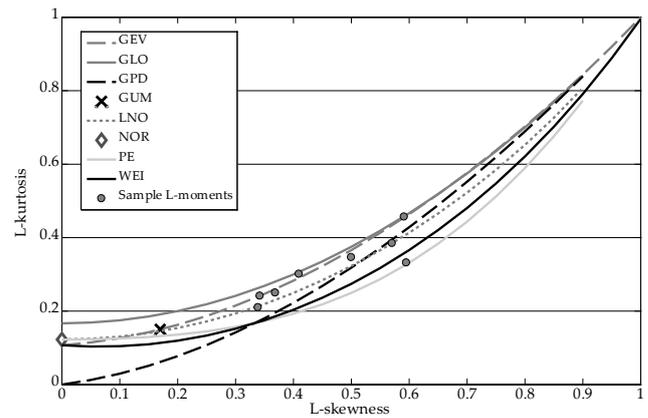
$$\lambda_1 = \xi + \lambda \{1 - \Gamma(1 + \kappa)\} / \kappa. \tag{14}$$

where  $\Gamma(\cdot)$  represents the gamma function.

The other parameter of the model is the  $p_{nz}$  value that evaluates the regime type of the basin and can range from 0 for basins with no flow in the period of measurements to 1 for rivers with permanent regimes.

#### 4.4.2 Regression models

Stepwise regression analysis was performed for all stations in the region. For this type of statistical regression model, the order of entry of the predictor variables is based on an F-test.



**Fig. 10.** L-moment ratio diagram of L-kurtosis vs. L-skewness, used to choose the parent distribution. Distributions in the diagram are: Normal (NOR), Gumbel (GUM), Generalized extreme values (GEV), Generalized logistic (GLO), Generalized Pareto (GPD), Generalized normal (LNO), Pearson type III (PE), Weibull (WEI).

**Table 8.** Observed parameters for the stations of Region 4 (Apennine area) used for the regionalisation approach.  $\mu$  and  $\sigma$  are the mean and standard deviation of the AF data;  $\kappa$ ,  $\lambda$  and  $\xi$  are the shape parameter, scale parameter and location parameter of the GEV distribution of the  $X'$  data;  $p_{nz}$  is the percentage of time that the river is flowing.

	$\mu$	$\sigma$	$\kappa$	$\lambda$	$\xi$	$p_{nz}$
Site 1	20.26	3.03	0.33	0.16	0.83	100.00
Site 2	6.44	1.38	0.29	0.24	0.77	100.00
Site 3	0.04	0.02	0.57	0.32	0.40	95.00
Site 4	5.73	3.89	2.25	0.16	0.07	100.00
Site 5	1.12	0.49	2.01	0.15	0.08	97.00
Site 6	8.95	1.47	0.35	0.22	0.76	100.00
Site 7	36.47	4.88	0.26	0.10	0.91	100.00
Site 8	76.94	25.39	0.51	0.29	0.58	100.00

A validation analysis, through jack-knife procedure, is then necessary to evaluate the accuracy of the regional estimates. This method allows a simulation of the presence of ungauged basins and is assessed using the procedure below (Castellarin et al., 2004).

One of the stations of the homogeneous region was removed from the sample and a new regression analysis was carried out without it. New parameters were then calculated for this station with the new equations; the results were used to calculate the FDC. The procedure was then applied to all the stations.

The stepwise procedure was then used to define regionalisation models to calculate the five parameters in ungauged basins. Three kinds of models (linear, exponential and logarithmic) were evaluated:

$$\hat{\vartheta} = A_0 + A_1\omega_1 + A_2\omega_2 + A_n\omega_n + \vartheta; \tag{15}$$

$$\hat{\vartheta} = A_0 \cdot \omega_1^{A_1} \cdot \omega_2^{A_2} \cdot \omega_n^{A_n} + \vartheta; \tag{16}$$

$$\hat{\vartheta} = A_0 + \ln(A_1 \cdot \omega_1) + \ln(A_2 \cdot \omega_2) + \ln(A_n \omega_n) + \vartheta. \tag{17}$$

where  $\hat{\vartheta}$  is the perfect estimated parameter,  $A_i, i = 1, 2, \dots, n$  are the coefficients of the model,  $\omega_i$  are the explanatory variables and  $\vartheta$  is the residual of the models.

The regression models identified for the five parameters of the modified stochastic index flow model are:

$$\mu = A_0 \cdot (A^{A_1}) \cdot (F_A^{A_2}); \tag{18}$$

$$\sigma = A_3 \cdot (A^{A_4}); \tag{19}$$

$$\kappa = A_5 \cdot (F_D^{A_6}) \cdot (H^{A_7}); \tag{20}$$

$$\lambda = A_8 \cdot (A^{A_9}); \tag{21}$$

$$p_{nz} = A_{10} \cdot (F_D^{A_{11}}) \cdot (A^{A_{12}}). \tag{22}$$

The two parameters of the AF data depend on the area of basin A, while parameters of the  $X'$  data and  $p_{nz}$  also depend on the percentage of pervious substrate  $F_D$  and on the percentage of calcareous substrate  $F_A$ .

Figure 11 shows the scatter plots of parameters observed versus the parameters predicted, which were calculated from all sites using the jack-knife procedure. It can be seen that the parameters estimated using the jack-knife procedure are more scattered than those obtained with general regression; however, they are not so different as to imply that the models cannot be used.

### 4.4.3 Results of regionalisation

The flow duration curves calculated using parameters obtained through jack-knife validation were then compared with empirical FDC in order to evaluate the accuracy of the model.

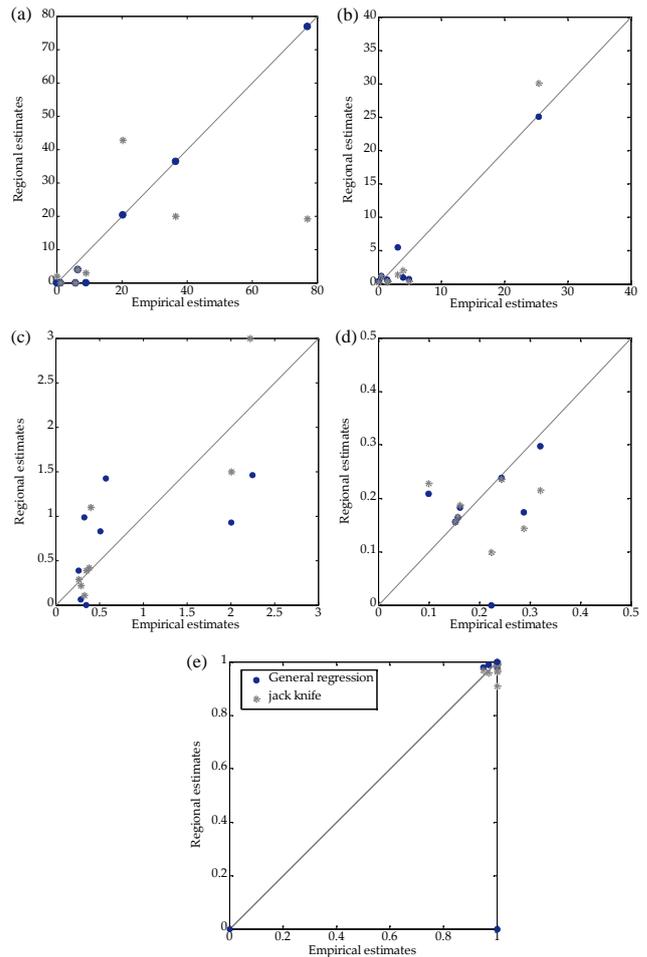
To evaluate the performance of the model, the following indicator was considered (Castellarin et al., 2004):

$$\varepsilon_{s,j} = \frac{\hat{q}_{s,j} - q_{s,j}}{q_{s,j}} \cdot 100 \tag{23}$$

where  $q_{s,j}$  and  $\hat{q}_{s,j}$  indicate the daily streamflow, empirical and estimated through regionalisation, associated with duration  $j$  for station  $s$ . From these values it is possible to obtain the mean relative error  $\bar{\varepsilon}_s$  and its standard deviation  $\sigma_{\varepsilon,s}$  for a station as:

$$\bar{\varepsilon}_s = \frac{1}{N'} \cdot \sum_{j=1}^{N'} \varepsilon_{s,j} \tag{24}$$

$$\sigma_{\varepsilon,s} = \sqrt{\frac{1}{N'} \cdot \sum_{j=1}^{N'} (\varepsilon_{s,j} - \bar{\varepsilon}_s)^2} \tag{25}$$



**Fig. 11.** Scatter plot of observed parameters vs. predicted parameters of the stochastic index flow model. Graphs (a) and (b) represent  $\mu$  and  $\sigma$  parameters of the normal distribution. Graphs (c) and (d) represent  $\kappa$  and  $\lambda$  parameters of the GEV distribution, and Graph (e) is the scatter plot of the  $p_{nz}$  values (in probability form).

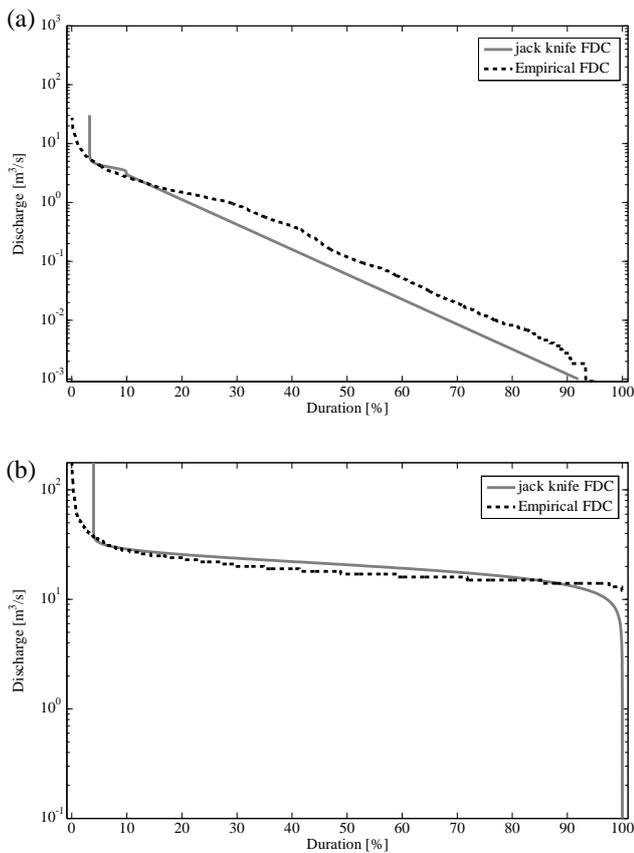
where  $N'$  represents the number of durations that are considered to calculate mean and standard deviation.

In addition, the average of  $N$  values of Eq. (24),  $\bar{\varepsilon}$ , and of Eq. (25),  $\sigma_{\varepsilon}$ , with  $N$  corresponding to the number of sites, gives an indication of the performance of the model.

It is also possible to graphically represent the mean and median of the distribution of the  $N$  relative errors  $\varepsilon_{i,j}$  and the  $100(\alpha/2)\%$  and  $100[1 - (\alpha/2)]\%$  percentiles, by identifying the interval about the median containing the  $100(1 - \alpha)\%$  of the  $N$  relative errors, against durations  $j$ , to evaluate the uncertainty of the regional FDC for all durations.

The mean relative error for a given duration  $j$  can be calculated as:

$$\bar{\varepsilon}_j = \frac{1}{N} \sum_{i=1}^N \varepsilon_{j,i} \tag{26}$$



**Fig. 12.** Semi-logarithmic representation of empirical and jack-knife FDC for site n. 5 (intermittent regime) (a) and for site n. 1 (permanent regime) (b).

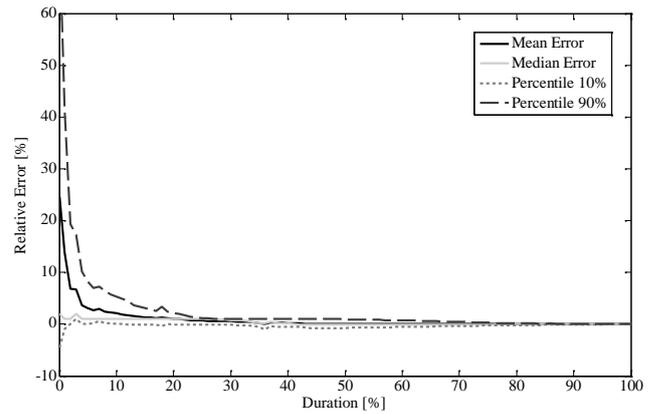
Another performance index  $E_s$  that can be used is the Nash-Sutcliffe efficiency method, calculated for each station as:

$$E_s = 1 - \frac{\sum_{j=1}^{N'} (\hat{q}_{s,j} - q_{s,j})^2}{\sum_{j=1}^{N'} (q_{s,j} - \bar{q}_s)^2} \quad (27)$$

where  $\hat{q}_{s,j}$  is the estimated value for each duration  $j$  and site  $s$ ,  $q_{s,j}$  is the empirical value for each duration  $j$  and site  $s$  and  $\bar{q}_s$  is the mean value. The value of this index can range between 1 and  $-\infty$ .

The  $E_s$  values are used to calculate three indexes of the effectiveness of the model:

- $P_1$  is defined as the percentage of cases over  $N$  stations in which  $E_s > 0.95$ ;
- $P_2$  is defined as the percentage of cases over  $N$  stations in which  $0.50 < E_s < 0.95$ ;
- $P_3$  is defined as the percentage of cases over  $N$  stations in which  $E_s < 0.5$ .



**Fig. 13.** Representation of mean (solid black line), median (solid grey line) and 10% and 90% percentiles (broken grey and black lines) of relative error for different durations.

**Table 9.** Indexes of reliability calculated on FDC obtained from the regionalisation model.

$\bar{\epsilon}_s$ [%]	0.96
$\sigma_\epsilon$ [%]	3.68
$P_1$ [%]	0.54
$P_2$ [%]	0.18
$P_3$ [%]	0.27

Figure 12 shows FDC results for two explanatory sites. Table 9 shows the results of the average relative error and the standard deviation of relative error for the region and the Nash-Sutcliffe efficiency index. Figure 13, on the other hand, is a graphic representation of mean, median and 10% and 90% percentiles of relative errors.

The relative error graph shows that the worst results are produced for shorter durations, but the error decreases; for durations greater than 20%, it is lower than 1%. The median value is less affected by the anomalous values and is always under 1%, very near 0%. It is particularly important to note that the relative error calculated for the higher percentage of durations, which coincides with the lower part of the FDC, is very low. The efficiency results are quite good: in fact, more than 50% of the sample is very well fitted, and only 27% fits poorly.

## 5 Summary and conclusions

The principal aim of this paper is to create a method for calculating FDC in the presence of zero data caused by zero flow in basins with intermittent regime, where environmental problems caused by wastewater are critical.

The model represents an improvement and a more general representation of the stochastic index flow model developed

by Castellarin et al. (2004), in that it utilizes the theorem of total probability. This is used to calculate the probability of all data  $\Pr(X \leq x)$ , while the stochastic index flow model is used to calculate the conditional distribution  $\Pr(X \leq x | X \neq 0)$ .

The method is introduced in a context of regionalisation; consequently, a procedure to transfer this model has been proposed. The starting point is the definition of homogeneous regions in the study area using cluster analysis and geomorphological variables as explanatory variables. Multiple regression analysis then brought about the definition of certain equations needed to transfer hydrological information to ungauged basins. At the end, an evaluation through jack-knife validation was performed to simulate the use of the model in the case of ungauged basins.

The definition of homogeneous regions was applied to basins in Central Italy and the modified stochastic index flow model was tested in a homogeneous region involving basins with intermittent flow.

Moreover, from the preliminary analysis of the discharge, four groups of nearby stations were identified. The factors with the most influence on the similarity of the data are the average altitude of the basin, the climate, the proximity of the observed stations to the coastline and the type of substrate. Cluster analysis using these geomorphological characteristics was then applied, and four corresponding regions were found according to the climatic condition of the area. The preliminary analysis showed a high heterogeneity in the area, caused by the orography that influences the climate and by the different types of substrate.

The results of the application of the five-parameter model show that it can be used to represent FDC on rivers with zero flow. In fact, the application of the modified model has produced a better approximation of FDC for one of the two stations analysed. Although a worse result is provided for the other one, this is probably caused by over-abstraction of groundwater. Furthermore, the little difference between the two models is due to the small percentage of zero data. Surely an application to a larger number of basins, characterized by a significant number of zero values, would be necessary to better evaluate the performance of this technique.

The regionalisation of the model and the jack-knife validation also provided good results. The implementation of the regionalisation model shows that the approach used based on the regression relationships and identified through step-wise regression relationships, can be adopted in different geographical areas by simply changing the explicative variables. It is important to underline that both the regionalisation approach and the validation were applied to a modest sample, since there was only one region in the study area with intermittent gauging stations. Only a few stations were thus used, of which only two had zero flow. Hence, further studies are necessary to test the model's applicability in regions with more stations with zero flow. Furthermore, the nested structure of the region influences the results of both

the homogeneity tests and the regression models. Therefore, further work should develop a procedure to regionalise the model in regions characterized by nested structures and create a model to compute annual-based FDC in intermittent rivers.

*Acknowledgements.* The authors thank anonymous reviewer and Andreas Efstratiadis for their helpful comments.

Edited by: R. Deidda

Reviewed by: A. Efstratiadis and another anonymous referee

## References

- Castellarin, A., Vogel, R. M., and Brath, A.: A stochastic index flow model of flow duration curves, *Water Resour. Res.*, 40, W03104, doi:10.1029/2003WR002524, 2004.
- Cigizoglu, H. K. and Bayazit, M.: A generalized seasonal model for flow duration curve, *Hydrol. Process.*, 14(6), 1053–1067, 2000.
- Claps, P. and Fiorentino, M.: Probabilistic flow duration curves for use in environmental planning and management, in: *Integrated Approach to Environmental Data Management Systems*, edited by: Harmancioglu, N. B., Alpaslan, M. N., Ozkul S. D., Singh, V. P., Kluwer Academy, NATO ASI Series, Ser. 2(31), 255–266, 1997.
- Dalrymple, T.: *Flood frequency analyses*, US Geological Survey, Water Supply Papers, 1543-A, 1960.
- Durrans, S. R., Ouarda, T. B. M. J., Rasmussen, P. F., and Bobée, B.: Treatment of zeroes in tail modelling of low flows, *J. Hydrol. Eng.*, 4(1), 19–27, doi:10.1061/(ASCE)1084-0699(1999)4:1(19), 1999.
- Ganora, D., Claps, P., Laio, F., and Viglione, A.: An approach to estimate nonparametric flow duration curves in ungauged basins, *Water Resour. Res.*, 45, W10418, doi:10.1029/2008WR007472, 2009.
- Gustard, A., Bullock, A., and Dixon, J. M.: *Low flow estimation in the United Kingdom*, Institute of Hydrology, Wallingford, UK, Report n. 108, 1992.
- Haan, C. T.: *Statistical methods in hydrology*, Iowa State University Press, Ames, Iowa, 1977.
- Hosking, J. R. M. and Wallis, J. R.: The effect of inter-site dependence on regional flood frequency analysis, *Water Resour. Res.*, 24(4), 588–600, doi:10.1029/WR024i004p00588, 1988.
- Hosking, J. R. M. and Wallis, J. R.: *Regional frequency analysis*, Cambridge University Press, New York, 1997.
- Jennings, M. E. and Benson, M. A.: Frequency curve for annual flood series with some zero events or incomplete data, *Water Resour. Res.*, 5(1), 276–280, doi:10.1029/WR005i001p00276, 1969.
- Kilmartin, R. F. and Peterson, J. R.: Rainfall-runoff regression with logarithmic transforms and zeros in the data, *Water Resour. Res.*, 8(4), 1096–1099, doi:10.1029/WR008i004p01096, 1972.
- Kroll, C. N. and Stedinger, J. R.: Estimation of moments and quantiles using censored data, *Water Resour. Res.*, 32(4), 1005–1012, 1996.
- Laaha, G. and Blöschl, G.: A national low flow estimation procedure for Austria, *Hydrological Sciences, Journal des Sciences Hydrologiques*, 52(4), 625–644, 2007.

- LeBoutillier, D. V. and Waylen, P. R.: A stochastic model of flow duration curves, *Water Resour. Res.*, 29(10), 3535–3541, doi:10.1029/93WR01409, 1993.
- Mimikou, M. and Kaemaki, S.: Regionalization of flow duration characteristics, *J. Hydrol.*, 82(1–2), 77–91, doi:10.1016/0022-1694(85)90048-4, 1985.
- Nikic, Z. and Radonja, P.: Modelling the influence of hydrogeological parameters on low flow in hilly and mountainous regions of Serbia, *Hydrolog. Sci. J.*, 54(3), 484–496, doi:10.1623/hysj.54.3.484, 2009.
- Quimpo, R. G., Alejandrino, A. A., and McNally T. A.: Regionalised flow duration curves for Philippines, *J. Water Res. Pl.-ASCE*, 109(4), 320–330, doi:10.1061/(ASCE)0733-9496(1983)109:4(320), 1983.
- Searcy, J. C.: *Manual of hydrology 2, Low flow techniques, flow duration curves*, US Geological Survey, Water Supply Paper, 1542-A, 1959.
- Smakhtin, V. U.: Low flow hydrology: a review, *J. Hydrol.*, 240, 147–186, 2001.
- Smakhtin, V. Y., Hughes, D. A., and Creuse-Naudine, E.: Regionalization of daily flow characteristics in part of the Eastern Cape, South Africa, *Hydrolog. Sci. J.*, 42(6), 919–936, 1997.
- Strupczewski, W. G., Węglarczyk, S., and Singh, V. P.: Impulse response of the kinematic diffusion model as a probability distribution of hydrologic samples with zero values, *J. Hydrol.*, 270(3–4), 328–35, doi:10.1016/S0022-1694(02)00309-8, 2003.
- Subcommittee on Hydrology: *Methods of Flow Frequency Analysis*, Inter-Agency Committee on Water Resources, Notes on hydrological act, Bulletin 13, Washington D.C., April 1966.
- Tate, E. L. and Freeman, S. N.: Three modelling approaches for seasonal droughts in southern Africa: the use of censored data, *Hydrolog. Sci. J.*, 45(1), 27–42, 2000.
- Vogel, R. M. and Fennessey, N. M.: Flow duration curves: new interpretation and confidence intervals, *J. Water Res. Pl.-ASCE*, 120(4), 485–504, doi:10.1061/(ASCE)0733-9496(1994)120:4(485), 1994.
- Wang, C. Y. and Singh, R. P.: Frequency estimation for hydrological samples with zero data, *J. Water Res. Pl.-ASCE*, 121(1), 98–108, doi:10.1061/(ASCE)0733-9496(1995)121:1(98), 1995.
- Ward J. H.: Hierarchical grouping to optimize an objective function, *J. Am. Stat. Assoc.*, 58(301), 236–244, doi:10.2307/2282967, 1963.
- Woo, M. K. and Wu, K.: Fitting annual floods with zero flows, *Can. Water Resour. J.*, 14(2), 10–16, 1989.
- Yu, P. S., Yang, T. C., and Wang, Y. C.: Uncertainty analysis of regional flow duration curves, *J. Water Res. Pl.-ASCE*, 128(6), 424–430, doi:10.1061/(ASCE)0733-9496(2002)128:6(424), 2002.