An approach to estimate nonparametric flow duration curves in ungauged basins

D. Ganora,[1] P. Claps,[1] F. Laio,[1] and A. Viglione[2]

Received 23 September 2008; revised 2 April 2009; accepted 8 July 2009; published 15 October 2009.

[1] A distance-based regionalization model is developed for the estimation of dimensionless flow duration curves (FDC) in sites with no or limited available data. The curves are dimensionless because they are preliminarily normalized by an index value (e.g., the mean annual runoff). The model aims at representing the FDC as a nonparametric object rather than providing a parametric representation and trying to relate the parameter values to basin descriptors. The regional approach considers the (dis)similarity between all possible pairs of curves and uses distance measures that can be related to basin descriptors, taken among geographic, geomorphologic, and climatic parameters. The (dis)similarity between curves is computed using a predefined metric based on a linear norm and produces a distance matrix. This matrix is then related, by means of linear regression models, to analogous matrices composed of the difference between all possible values of each descriptor within the set of basins. After identification of significant descriptors, a cluster analysis is applied so that the basins can be grouped together. Each region is supposed to be characterized by a single dimensionless flow duration curve. The procedure is applied to 95 basins located in northwestern Italy and Switzerland. The performance in the regional estimation is assessed by means of a cross-validation procedure through comparison with "standard" parametric regional approaches based on two- and three-parameter models. In most of the cases, the distance-based model produces better estimates of the flow duration curves using only few catchment descriptors.

Citation: Ganora, D., P. Claps, F. Laio, and A. Viglione (2009), An approach to estimate nonparametric flow duration curves in ungauged basins, *Water Resour. Res.*, *45*, W10418, doi:10.1029/2008WR007472.

1. Introduction

[2] The problem of estimating hydrological variables in ungauged basins has been the object of intense research activity in recent years [see, e.g., *Sivapalan et al.*, 2003]. Regardless of the method used to perform such estimation, the underlying idea is to transfer the hydrological information from gauged to ungauged sites. When observations of the same variable at different measuring sites are available and are used for the estimation in ungauged sites, the related methods are called regional methods. Regional frequency analysis [e.g., *Hosking and Wallis*, 1997], where the interest is in the assessment of the frequency of occurrence of hydrological events, belongs to this class of methods. A frequently used regional approach is the index value method [*Dalrymple*, 1960] in which it is implicitly assumed that the frequency distribution for different sites belonging to a homogeneous region is the same except for a site-specific scale factor, the index value (see, e.g., *Hosking and Wallis* [1997] for details). Hence, the estimation of the distribution for an ungauged site is obtained by separately estimating the index value and assigning the site to an homogeneous region, which entails assuming a dimensionless frequency distribution (a growth curve) to the site under analysis.

[3] The frequency distribution is only one of the possible information that can be transferred using a regional approach. In this paper we deal with a specific descriptor of the runoff distribution in a basin: the flow duration curve (FDC). A flow duration curve represents the flow in a stream rearranged to show the percentage of time during which a discharge value is equaled or exceeded. Strictly speaking this is not a probability curve, because discharge is correlated between successive time intervals and discharge characteristics are dependent on the season; hence the probability that discharge on a particular day exceeds a specified value depends on the discharge on preceding days and on the time of the year [*Mosley and McKerchar*, 1993, p. 8.27]. However, a flow duration curve is often interpreted as the complement of the cumulative distribution function of the daily streamflow values at a site. The FDC also provides a graphical summary of streamflow variability and is often used in hydrologic studies for hydropower, water supply, irrigation planning and design, and water quality management (a review on many applications is provided by *Smakhtin* [2001]).

[4] The empirical FDC is constructed from observed streamflow time series. These observations can have different time scale resolution, although mean daily streamflow

[1]Dipartimento di Idraulica, Trasporti ed Infrastrutture Civili, Politecnico di Torino, Turin, Italy.
[2]Institut für Wasserbau und Ingenieurhydrologie, Technische Universität, Vienna, Austria.

0043-1397/09/2008WR007472$09.00

values are commonly used. The data are ranked in descending order and each ordered value is associated with an exceedance probability F, for example through a plotting position formula. If the FDC is constructed on the basis of the whole available data set, merging together all available years of data, it represents the variability of flow over the entire observation period. This representation is valid when the data set is sufficiently long. A different approach, introduced by *Vogel and Fennessey* [1994], is to consider annual FDCs separately, i.e., to consider a different FDC for each year when data are available [e.g., *Claps and Fiorentino*, 1997; *Iacobellis*, 2008]. A parametric model able to represent both the total and the annual FDCs for gauged and ungauged sites has been proposed, for instance, by *Castellarin et al.* [2004b, 2007].

[5] In the present work only total FDCs will be considered, adopting a nonparametric approach for their representation. The FDCs are modelled following the index value approach, in which the flow duration curve $Q(F)$ is the product of two terms $Q(F) = \mu\, q(F)$, where the index flow μ is the scale factor and the dimensionless total flow duration curve $q(F)$ represents the shape of the FDC. The present work focuses on the regionalization of the dimensionless curve, while the estimation of the index flow will not be treated. In section 2 we describe the distance-based method. This method is applied to a case study in section 3, where a set of basins located in northwestern Italy and Switzerland is investigated. The method's performances against alternative parametric methods are finally checked in section 4.

2. Distance-Based Method

[6] Leaving aside the index flow estimation, the regional FDC analysis can be divided into two parts: the formation of cluster regions and the association of an ungauged site to one of them. Concerning the first point, the curves are grouped according to their similarity in terms of the basin descriptors that better "explain" the shape of the FDC. In standard approaches [e.g., *Fennessey and Vogel*, 1990; *Singh et al.*, 2001; *Holmes et al.*, 2002], this shape is represented in a parametric way. For instance, the coefficient of variation (CV) or the L-CV [*Hosking and Wallis*, 1997] of the curve can be used for this purpose. In this case, the selected parameter is related to basin descriptors through a linear or a more complex model. A regression analysis is performed with different combinations of descriptors, and those that are strongly related with the parameter are used for its estimation in ungauged sites.

[7] In the distance-based approach proposed here we consider the dimensionless FDC as a whole, without resorting to statistical descriptors of its shape. This means that a curve is not fitted by an analytical function, which would imply a parametric representation of the FDC. The multiregression approach can still be used to study the (dis)similarity between pairs of basins. The procedure is synthetically described below as a sequence of logical steps, while details are provided in the following subsections: (1) for each couple of stations, a dissimilarity index between dimensionless curves is calculated using a predefined metric (section 2.1); (2) for each considered basin descriptor (e.g., area, mean elevation, mean slope, drainage path length, etc), the absolute value of the difference between its measure in two basins is used as the descriptor distance; (3) the distances between couples of FDCs (and between basin descriptors) are organized in distance matrices (section 2.2); (4) a multiregression approach is applied using the FDC distance matrix as the dependent variable, and the descriptor distance matrices as the independent variables; this serve to select the relevant basin descriptors (those associated to the best regression model) (section 2.2); (5) in the resulting descriptors' space, stations with similar descriptor values (small distances between descriptors) are grouped together into regions through a cluster analysis (section 2.3); and (6) the regional dimensionless flow duration curve is estimated by taking the average of all the curves belonging to the cluster, as in the "graphical approaches" reviewed by *Castellarin et al.* [2004a, and references therein].

[8] Critical points of this procedure, discussed more in detail in the following, are the choice of a suitable distance measure for the dimensionless flow duration curves, the identification of the best regression model between distance matrices, and the choice of the method of cluster analysis for the formation of the regions.

2.1. (Dis)similarity Between Curves

[9] Let Q^*_s be the sequence of N_s daily discharges in the gauged station s, containing all the recorded values. Based on these data the scale factor μ_s is first computed as the average of the whole sequence. Then, the dimensionless sequence $q^*_s = Q^*_s/\mu_s$ is rearranged in descending order and each value $q_{i,s}$, with $i = 1, 2, \ldots, N_s$, is associated to its exceedance probability (i.e., through the Weibull plotting position)

$$\left\{\frac{1}{N_s+1}, \frac{2}{N_s+1}, \ldots, \frac{N_s}{N_s+1}\right\}. \tag{1}$$

The distance-based procedure proposed here is based on the comparison between couples of curves: for this purpose it is convenient the two curves have the same number of elements. Since total FDCs have generally different lengths, depending on the number of years they cover, we resample them to make the curves comparable. For this purpose, we resample the FDCs at the frequency values

$$\left\{\frac{1}{365+1}, \frac{2}{365+1}, \ldots, \frac{365}{365+1}\right\}, \tag{2}$$

obtaining a new representation of the FDC in the station s, $\{q_{1,s}, q_{2,s}, \ldots, q_{365,s}\}$. Other sampling rates can be used to better sample particular parts of the curves. In this work we have also considered an alternative sampling method that produces 365 equally spaced values in the z space, where z is the normal reduced variate (with zero mean and unit variance). Back transforming these values to the frequency space, the 365 values are no more equally spaced but more concentrated around higher and lower frequencies. Figure 1 sketches two curves with different number of elements resampled with a constant and a z spacing in the frequency axis.

[10] In our approach a measure of similarity between curves (hereafter termed distance) is required. Given two FDCs, relative to two gauging stations s_1 and s_2, constituted by 365 elements each: $\{q_{1,s1}, q_{2,s1}, \ldots, q_{365,s1}\}$ and $\{q_{1,s2}, q_{2,s2}, \ldots, q_{365,s2}\}$, a simple measure of their dissimilarity

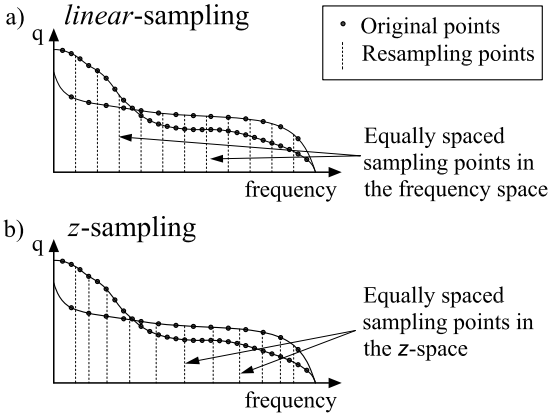

Figure 1. Comparison of dimensionless flow duration curves: sampling points (a) with constant spacing in frequency representation and (b) with a denser presence on the FDC tails due to normal transformation.

can be defined as the "distance" calculated by the norm of order 1,

$$\delta_{s_1,s_2} = \sum_{i=1}^{365} |q_{i,s_1} - q_{i,s_2}|. \quad (3)$$

The value $\delta_{s1,s2}$ can be interpreted also as an approximation of the area between the curves. The computation of the distance according to equation (3) is exemplified in Figure 2 for two generic FDCs.

[11] If n is the number of sites where data are available, the distance measures for each FDC pair are organized in a $n \times n$ distance matrix like

$$\Delta = \begin{pmatrix} 0 & \delta_{1,2} & \dots & \delta_{1,n} \\ \delta_{2,1} & 0 & & \vdots \\ \vdots & & \ddots & \\ \delta_{n,1} & \dots & & 0 \end{pmatrix}, \quad (4)$$

where the elements $\delta_{s1,s2}$ are distances between curves (calculated with equation (3)). Analogously, matrices like (4) can contain distances between catchment descriptors (if d_1 is the value of the descriptor for basin 1 and d_2 for basin 2, then $\delta_{1,2} = |d_1 - d_2|$). Since the matrices are symmetric and with null diagonal values, after removing the redundant values, only $n(n-1)/2$ values per matrix are informative.

[12] The distance measure of equation (3) not only depends on the resampling method but also on the "measurement space" considered for the representation of flows. For example, if the flows are transformed to provide a more convenient representation of the FDC, the distances $\delta_{s1,s2}$ are affected by the transformation. Three main representations of the FDC are considered in this work: (1) flow data plotted versus their corresponding plotting position, (2) log-transformed flows versus their corresponding plotting position and (3) lognormal probability plot (log-transformed flows versus normal reduced variate). There are no particular reasons to prefer a priori one of these representations, therefore all of them are considered in the case study and will be referred to as "linear representation", "logarithmic representation" and "lognormal representation," respectively (see Figure 2). Three parametric functions will be used in a traditional regional FDC estimation exercise in section 4, for comparison to the distance-based procedure developed here.

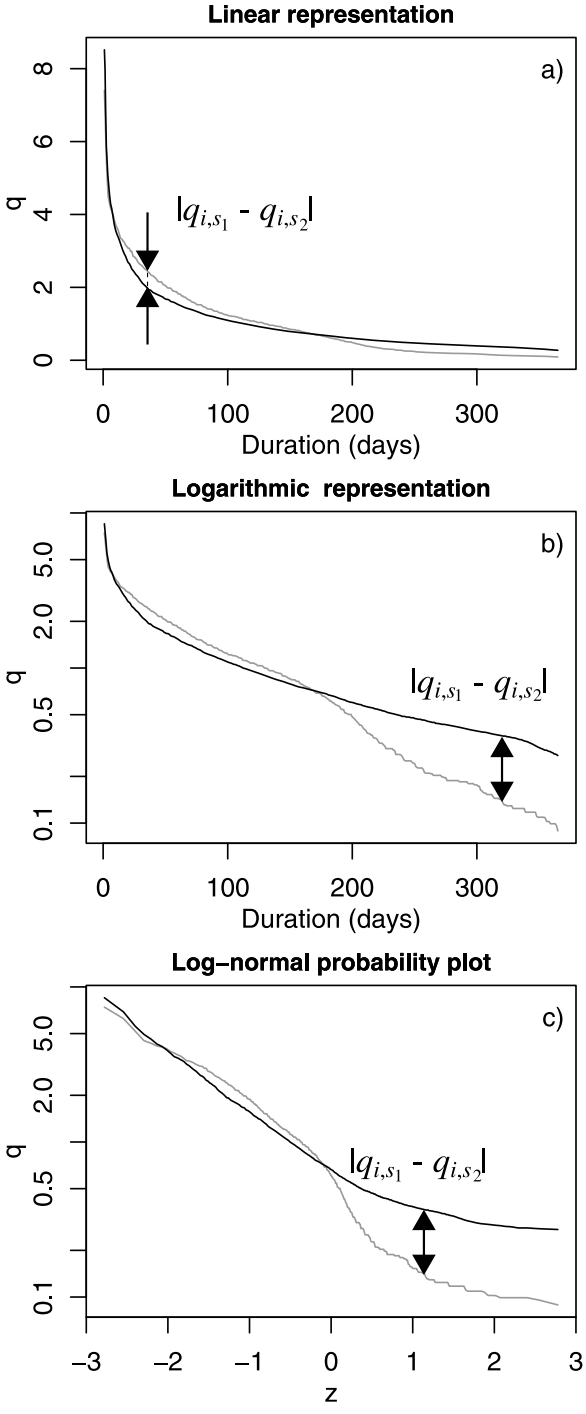

Figure 2. Distance between two FDCs calculated following equation (3). Shown are a pair of FDCs in three different representation spaces: (a) the linear representation (flow values versus exceedance frequency), (b) the logarithmic representation in which discharges are log transformed, and (c) the lognormal probability plot in which the abscissa is the normal reduced variate z.

2.2. Distance Matrices, Linear Regression, and Mantel Test

[13] In this section we show how to identify the catchment descriptors that, thanks to their relations with the FDCs, should be used for the formation of cluster regions. A different distance matrix, hereafter $\mathbf{\Delta}_{Xi}$, is determined for each descriptor, while the distance matrix for the dimensionless FDCs is called $\mathbf{\Delta}_Y$. The relation between the distance matrix $\mathbf{\Delta}_Y$ and the various $\mathbf{\Delta}_{Xi}$ is assessed using a multiregressive approach. Note that the multiregressive approach based on distance matrices is not used to estimate FDC coefficients, but to identify the descriptors to be used in the following step for region creation. We start considering a simple linear model:

$$\mathbf{\Delta}_Y = \beta_0 + \beta_1 \mathbf{\Delta}_{X_1} + \ldots + \beta_p \mathbf{\Delta}_{X_p} + \varepsilon \tag{5}$$

with p as the number of descriptors involved, β_i as the regression coefficients and ε the residual matrix. The best possible regression is selected through the adjusted coefficient of determination

$$R^2_{adj} = 1 - \left(1 - R^2\right) \frac{n-1}{n-p-1}, \tag{6}$$

where R^2 is the standard coefficient of determination [e.g., *Kottegoda and Rosso*, 1997], p the number of descriptors and n the number of basins considered. The regression coefficients and R^2 can be computed in a standard way [*Legendre et al.*, 1994], that is to say that it does not matter if the elements are organized in a distance matrix. However, in the formulation of the adjusted coefficient of determination it is better to use the value n (the number of basins) instead of $n(n-1)/2$ that is the number of points involved in the regression (namely the number of distance values). This is due to the fact that the values inside the matrices are not mutually independent. Dependancy has another significant impact on the method. In particular, the validity of the tests used to assess the significance of the independent variables (e.g., the Student t test) is affected. A different significance test, as the Mantel test [*Mantel and Valand*, 1970], is then needed, which accounts for the nonindependence of the elements in the distance matrices.

[14] The Mantel test was originally proposed by *Mantel and Valand* [1970] for analysis of correlation between distance matrices, and since then it has been widely improved and used with many different kinds of data. In fact, distance matrices have been frequently used in the biological and ecological sciences [e.g., *Legendre*, 1993; *Lichstein*, 2007]. The simple Mantel test [*Mantel and Valand*, 1970] is used to evaluate the significance of the linear correlation between two distance matrices. This test is performed computing a statistic (usually the Pearson correlation coefficient) between all the pairwise elements of the two matrices. Its significance is tested by repeatedly permuting the objects in one of the matrices, and recomputing the correlation coefficient each time; permutations are performed simultaneously exchanging the rows and the columns of the matrices (e.g., if rows of indexes 2 and 10 are exchanged, also columns of indexes 2 and 10 have to be exchanged (see *Legendre et al.*, 1994)). The significance of the statistic is assessed by comparing its original value to the distribution of values obtained from the permutations, which are considered as many realizations of the null hypothesis of no correlation.

[15] The simple Mantel test can be extended to multiple predictor variables to be applied in multiple linear regression models as (5). The extension has been introduced by *Smouse et al.* [1986], discussed and improved by *Legendre et al.* [1994] and recently applied in the ecological field by *Lichstein* [2007]. Following the procedure of *Lichstein* [2007] each matrix, after removing redundant values, is unfolded into a vector of distances, and regression is performed in the classical way. Then, a null distribution is constructed permuting the elements only in the dependent variable distance matrix $\mathbf{\Delta}_Y$. Similarly to what described for the simple Mantel test, the rows and the columns of the matrix $\mathbf{\Delta}_Y$ are permuted simultaneously and each regression coefficient is tested individually.

2.3. Cluster Analysis

[16] The proposed procedure serves for the estimation of a FDC in an ungauged basin on the basis of curves relative to other basins. Given a large group of candidate "donor" basins, we want to extract a subset of basins that have geomorphologic and climatic characteristics similar to those of the target site. The FDCs collected in these sites will be used for the estimation of the unknown curve. There are different regionalization techniques to choose the subset of basins, for example leading to the formation of fixed regions through cluster analysis [*Hosking and Wallis*, 1997; *Viglione et al.*, 2007b], or based on the method of the region of influence (ROI) [*Burn*, 1990]. In this work we use the first approach, selecting fixed regions by splitting the descriptors space in nonoverlapping areas by means of a cluster analysis. However, the generalization of the method to the ROI technique is straightforward. The definition of the descriptors space depends on the outcome of the multiregressive procedure described in section 2.2, that allows one to identify a group of significant geomorphoclimatic parameters.

[17] The cluster analysis method used here is a mixed method in which the Ward hierarchical algorithm [*Ward*, 1963] is followed by a reallocation procedure that minimizes the dispersion within each cluster. The Ward algorithm is agglomerative; it starts with a configuration in which each element is a cluster itself, and progressively merges clusters in a way to produce the minimum information loss, measured as the sum of squared deviation of each element from its cluster centroid. We use the Ward algorithm because it is able to generate compact clusters with an evenly distributed number of elements. A disadvantage is that it does not allow elements reallocation, so that the final configuration could not be the optimal one. To avoid this inconvenience, a reallocation procedure is applied in concurrence with the agglomerative clustering. For instance, if the Ward clustering yields a final configuration with k clusters we compute the statistic

$$W = \sum_{i=1}^{k} \left(\sum_{j=1}^{n_i} D^2_{i,j} \right), \tag{7}$$

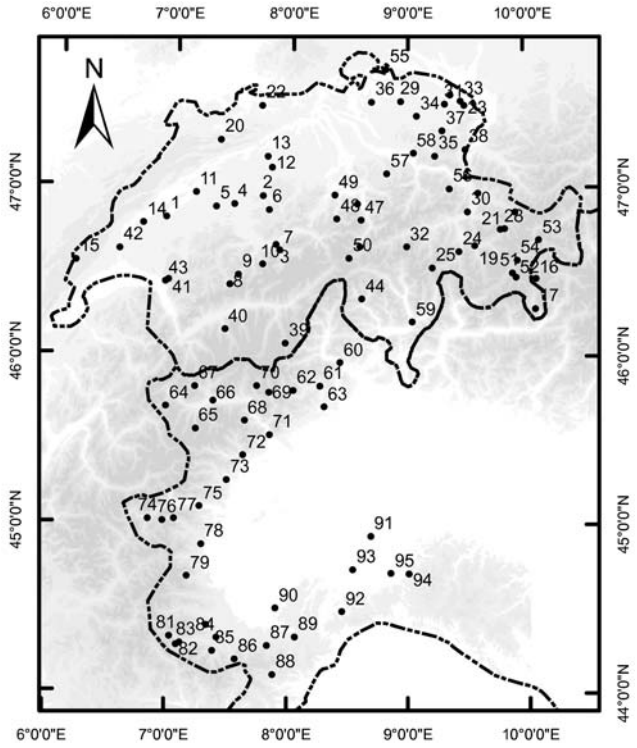

Figure 3. Geographical location of the gauging stations of the 95 catchments considered in the study. Basins 1–59 belong to Switzerland, while the remaining ones are located in the northwestern part of Italy, in the Piemonte and Valle d'Aosta regions. For additional information see http://www.idrologia.polito.it/~ganora.

where $D_{i,j}$ is the Euclidean distance between the jth element of the ith cluster and the cluster centroid, and n_i is the number of elements contained in the ith cluster. An element is moved to another cluster if the new configuration provides a lower value of W. The procedure ends when W stops reducing after the reallocations, so that every element of a cluster is closer to its center of mass than to the centroid of the nearby cluster.

[18] The reallocation procedure leads to an optimal configuration with k regions. A controversial point of the procedure is the choice of the optimal number of clusters. Usually, in regional analyses, the aim is to get the smallest possible number of homogeneous regions, so that each of them has a large enough number of elements. In this work, the selection of the ideal number of clusters is done investigating different k values and evaluating, for each configuration, a quality index. This index is computed by estimating (in cross-validation mode) the curves for all sites by using the regional model (with a given k), and computing the distance as in equation (3) where, in this case, s_1 is the measured curve and s_2 is the estimated one. This distance is adopted as an error measure and the overall mean error is used as a quality index to select the number of clusters. This method does not ensure that the clusters are homogeneous, because no homogeneity test is explicitly used.

[19] After having subdivided the descriptors space in regions, one can proceed to the estimation of the flow duration curve in ungauged sites. For one such site one must first determine the values of the descriptors selected in the procedure of section 2.2. The descriptors at the ungauged site are entered as coordinates in the descriptors' space and the site is assigned to the cluster whose centroid is the closest to the basin descriptors. The curves of all basins belonging to the selected cluster will be used to build the regional curve. This latter curve is simply estimated point by point as the average of the values of q relative to each duration for the curves belonging to the selected region, as in the graphical approach described by *Castellarin et al.* [2004a].

[20] The descriptors used in the cluster analysis are preliminary standardized (i.e., converted into variables with zero mean and unit variance). Standardization of raw descriptors values avoids an unwanted weighting effect due to the different measurement units. If the descriptors are assumed to have different importance in the cluster creation, a procedure can be adopted to give different weights to each descriptor. Regression coefficients of equation (5) can be used to compare the relative effect of each descriptor distance matrix, if the distance matrices have been previously standardized: the greater the coefficient, the greater the relative effect of its descriptor distance matrix on the curve distance matrix, so that the coefficients can be used as weights. This weighted clustering procedure will be tested in the following sections by comparing it to the standard unweighted clustering.

[21] After the regional curves have been determined, it is necessary to evaluate if they can be considered significantly different from each other, because otherwise the regions should be merged. To assess if two regional curves are significantly different, we use a procedure based on the distances between curves. First, a reference distance is computed as the median (or the mean) of the distances between each empirical curve and the regional one. Then, the distance matrix of the regional curves is computed and all its elements are compared against the reference distance: two regional curves are considered significantly different if their distance is greater than the reference distance, otherwise the two clusters are merged together. This procedure is repeated until all the regional FDCs are significantly different.

[22] Note that the reference distance and the distance matrix of the regional curves depend on the representation space on which the distances are calculated, hence different results are expected using different representation spaces.

3. Case Study: Distance-Based Method Application

3.1. Hydrological and Geomorphologic Data

[23] The application of the distance-based procedure for regional estimation of FDC has been carried out in the R statistical environment [*R Development Core Team*, 2007], integrated for Mantel test and cluster analysis with the nsRFA package [*Viglione*, 2007].

[24] Available data include 95 river basins located in northwestern Italy (36 basins of Piemonte and Valle d'Aosta regions) and in Switzerland (59 basins); the geographical location of the gauging stations is shown in Figure 3. Italian flow data derive from the publications of the former Italian Hydrographic Service and include series lengths ranging between 7 years and 41 years. Hydrological and geomor-

Table 1. Regression Models With Two Descriptors That Well Describe the Relationship Between Curve Distance Matrix and Descriptors Distance Matrices[a]

	Representation Space		
Best Relation	Linear	Logarithmic	Lognormal
First	H + MHL	Hmin + MHL	Hmin + MHL
Second	Hmin + MHL	Hmin + Pm	H + MHL
Third	H + Slo	Pm + MHL	Pm + MHL

[a]All the models pass the Mantel test (significance of regression coefficients) with a level of significance of 0.05 and the VIF test (multicollinearity) with threshold equal to 5. The curve distance matrix is calculated in three different representation spaces: the linear, the logarithmic, and the lognormal.

phological variables relative to Italian basins are included in the widest CUBIST database [*CUBIST Team*, 2007] that contain such data for more than 500 basins in Italy. The catchment area of northwestern Italy basins ranges between 22 and 7983 km^2, and their average elevation ranges from 494 to 2694 m a.s.l. Switzerland data are included in the Reference Hydrometric Network (SHRN) provided by the BAFU (Bundesamtes für UmweltSwiss) and include daily streamflow series with a minimum length of 18 years and a maximum length of 99 years. The catchment area of Switzerland basins ranges between 7 and 616 km^2, while their average elevation varies from 475 to 2847 m a.s.l. Geomorphological characteristics of each basin has been obtained from a digital terrain model (about 90 m cell grid) provided by *NASA* [2000] with automatic procedures originally developed by *Rigon and Zanotti* [2002] under a GRASS GIS environment. For the complete list of basins considered, whose codes are referred in Figure 3, and their geomorphologic variables see http://www.idrologia.polito.it/∼ganora.

3.2. Procedure Setting

[25] Several linear regression models between distance matrices have been investigated using relation (5). They are built using different combination of (1) curve distance matrices $\mathbf{\Delta}_Y$: the three representations described in section 2.1 and Figure 2 (linear, logarithmic and lognormal plot) are considered and (2) descriptors distance matrices $\mathbf{\Delta}_X$: all possible combination from one to five descriptors have been taken into account.

[26] Regression models are ordered in terms of R^2_{adj} values and tested for significance with the multiple Mantel test, with a significance level of 0.05. Furthermore, a test against multicollinearity has been performed in order to exclude variables with redundant information [*Montgomery et al.*, 2001].

[27] For the linear representation, best results are obtained with four and three descriptors. Lower R^2_{adj} values arise from simpler models with only two descriptors. In the logarithmic space, the best model is again characterized by four descriptors, but in this case simpler models with two parameters have comparable R^2_{adj}. In the lognormal space none of the solutions accepted after testing are based on more than two descriptors. We decided to adopt models with two parameters because of their higher robustness (see Table 1). The R^2_{adj} values obtained with regression models with distance matrices are very low, although the descriptors result to be statistically significant. In this regard it is important to remind that regressions are only used for the selection of the suitable descriptors and not for direct estimation.

[28] Table 1 shows the three best models for each representation with two descriptors, where all the models have been tested for significance of regression coefficients with the Mantel test with a level of significance of 0.05. It appears that, considering together the three representations of different curve distance matrices, the most significant descriptors are always the same: the minimum basin elevation (Hmin), the mean elevation (H), the mean hillslope length (MHL), the mean basin slope (Slo) and the modified basin slope (Pm, which is the ratio between the median elevation and the square root of the area). A summary of the range of these descriptors is reported in Table 2. This suggests to adopt the same set of descriptors with all the three representation spaces; Hmin and MHL has been selected. The adoption of these two descriptors is coherent with the typology of investigated basins. In fact, since we are considering mainly mountain basins, the elevation descriptor is expected to be relevant because of its strong relation to snow accumulation and snowmelt mechanisms. Similarly, the hillslope mean length provides a synthetic description of runoff routing mechanisms.

3.3. Regions Definition

[29] The second step, after the choice of the suitable descriptors, is to pool the catchments together with the cluster analysis, as described in section 2.3. The procedure is applied to both the weighted and the unweighed cluster configurations. For all the three representation spaces, the unweighed procedure often demonstrates better performances, while the weighted procedure leads to marginal, if any, improvements that do not justify its use. Following the criteria mentioned in section 2.3 and considering all the three representation spaces, the suggested number of clusters obtained for Italian and Switzerland data is four.

[30] This configuration is then checked, to assess if the regional FDCs are significantly different, using the procedure described in section 2.3 for all the three representation

Table 2. Brief Description and Range of Variation of the Descriptors Used by the Distance-Based Models[a]

Descriptor	Definition	Minimum	Mean	Maximum
H	mean elevation of the drainage basin above sea level (m)	475	1665	2847
Hmin	minimum elevation of the drainage basin above sea level (m)	82	839	1974
MHL	mean hillslope length (m)	584.1	759.5	973.6
Slo	average of the slope values associated to each pixel in the DEM of the drainage basin (%)	4	39.9	61.6
Pm	mean large-scale slope (%)	0.8	15.7	50.1

[a]See Table 1.

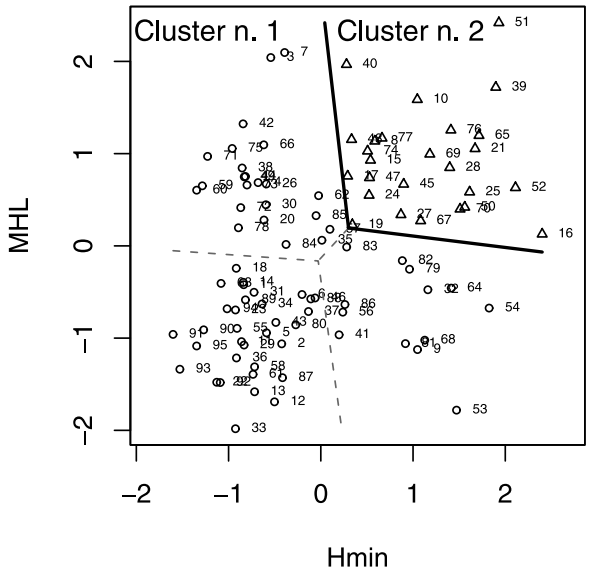

Figure 4. Disjoint regions in the space of catchment descriptors: Hmin is the minimum basin elevation and MHL is the mean hillslope length. The dashed lines represent the boundaries between the four clusters obtained before merging the clusters whose FDCs cannot be considered significantly different. The final two disjoint regions are separated by the solid line.

spaces. The FDCs of the original four clusters cannot be considered significantly different from each other, neither in the linear space, nor in the other two logarithmic spaces. Thus, for each representation space, the two most similar clusters are merged together. The new configurations with three clusters can be accepted in the linear space only. Applying again the procedure for the logarithmic and lognormal space we obtain two configurations consisting of two clusters each. To select one among these different configurations of clusters, we perform the following cross evaluation: for each set of clusters (e.g., the one obtained in the linear space), we check if the difference between the regional FDCs is significant in the other representation spaces (i.e., also in the logarithmic and lognormal spaces). Based on this cross evaluation, we choose the configuration with 2 clusters obtained in the logarithmic space, which is represented in Figures 4 and 5. Hence, this latter configuration will be used as the result of the distance-based model.

[31] The final regions obtained are shown in Figure 4. Curves belonging to each cluster are grouped together and the regional curves are derived as the average of all curves belonging to the region. Figure 5 shows the regional curves (black lines) obtained from curves belonging to the cluster (grey lines) in the lognormal space. Although every curve bundle appears to be quite wide, regional curves are able to represent two characteristic behaviors. In fact, we can observe an almost straight curve and a "S" shaped curve. A quantitative representation of model quality and estimation errors is reported in section 4, where a comparison against some parametric methods is performed.

4. Comparison With Parametric Models

[32] The distance-based regional procedure developed in this work is tested against some standard parametric regional models. In general, the choice of the reference model is not trivial and more than one function can be used to describe the FDCs. For this purpose, a useful tool is the L moments ratio diagram of Figure 6 [*Hosking and Wallis*, 1997] where one plots the L_{CA} (coefficient of L skewness) of each dimensionless FDC versus its corresponding L_{kur} (coefficient of L kurtosis). The lines represent the domain of the distributions over the $L_{CA} - L_{kur}$ space and can help one to identify the distribution to be used. This approach has been followed, for example, by *Castellarin et al.* [2007].

[33] In this work, the analysis is performed over a database of 95 basins that have very different characteristics in terms of L_{CA} and L_{kur}, as Figure 6 shows. The scattering of the points make the choice of the distribution rather difficult. For this reason, different parametric models are used for the comparison with the distance-based procedure.

[34] Each parameter θ of a parametric model is related to the catchments' descriptors d by a linear model of the form

$$\theta = a_0 + a_1 \cdot d_1 + a_2 \cdot d_2 + \ldots + a_n \cdot d_n + \varepsilon. \quad (8)$$

The first step is to identify a suitable regional model to estimate the generic parameter for an ungauged station, where θ is previously estimated at each station s using a suitable technique. The resulting parameters θ_s are then related to descriptor data (raw data, not distances) for all the catchments (not classified in regions) to identify a regional model (regression) able to describe them. Many linear

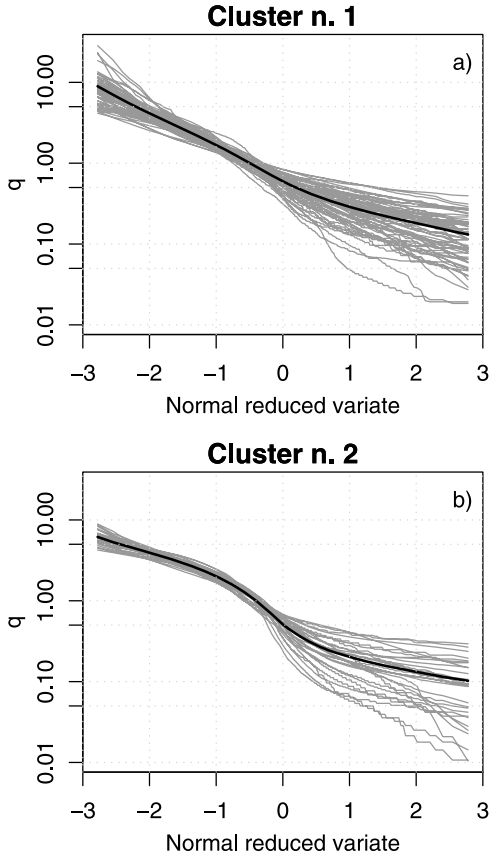

Figure 5. Flow duration curves grouped by cluster (grey) and corresponding regional curves (black).

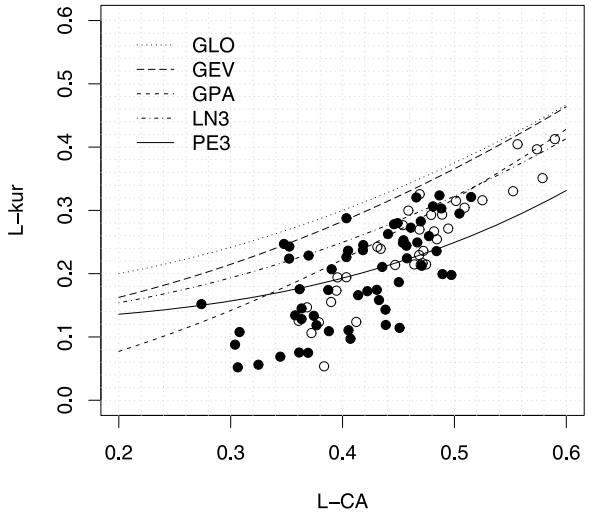

Figure 6. L moments ratio diagram for the dimensionless FDCs of the 95 basins (solid circles for Switzerland data and open circles for Italian data). The lines indicate different theoretical three-parameter distributions: generalized logistic (GLO), generalized extreme value (GEV), generalized Pareto (GPA), lognormal (LN3), and Pearson type III (PE3).

models of the form of equation (8) are considered and validated with a Student t test followed by a multicollinearity (VIF) test and subsequently ordered by their values of R^2_{adj} [e.g., *Montgomery et al.*, 2001].

[35] The models considered here are the two-parameter lognormal distribution (LN2), the three-parameter Pearson type III (PE3) and the generalized Pareto (GPA) distributions. The lognormal model is represented by the relation

$$\log(q) = \theta_1 + \theta_2 \cdot z, \tag{9}$$

where z is the quantile of a normal distribution with zero mean and unit variance corresponding to each flow's plotting position values. In the lognormal probability representation, equation (9) is a straight line whose coefficients θ_1 and θ_2 can be estimated with a least squares linear regression.

[36] The GPA probability density function is defined as

$$f(q) = \theta_2^{-1} \exp[-(1-\theta_3)y], \tag{10}$$

with $y = -\theta_3^{-1} \log[1 - \theta_3(q-\theta_1)/\theta_2]$ if $\theta_3 \neq 0$ and $y = (q - \theta_1)/\theta_2$ if $\theta_3 = 0$, where θ_1, θ_2 and θ_3 are the location, scale and shape parameter, respectively; the PE3 probability density function is defined as

$$f(q) = \frac{(q-\theta_1)^{\theta_2 - 1} \exp[-(q-\theta_1)/\theta_3]}{\theta_3^{\theta_2}\Gamma(\theta_2)}, \tag{11}$$

where θ_1, θ_2 and θ_3 are the location, scale and shape parameter, respectively, and $\Gamma(\cdot)$ is the gamma function. For details about these distributions and for parameters estimation we refer to *Hosking and Wallis* [1997] and *Viglione* [2007]. The regional estimation of the models' parameters use the descriptors listed in Table 3, whose definitions [*Viglione et al.*, 2007a] are available at http://www.idrologia.polito.it/~ganora.

[37] Our model and the parametric ones are all tested using a cross-validation approach in which one station is considered ungauged and its data are removed from the database. The models are then recalibrated using only the remaining data, and the unknown curve is estimated. After this procedure is repeated for all basins, one can compute, for each basin, the error measure $\delta_{MOD,EMP}$ as the distance between the estimated FDC and its empirical counterpart.

[38] The nonparametric FDC representation method performs better than the parametric models for most of the analyzed basins, independently of the representation space considered. Figure 7 shows a comparison between the errors $\delta_{MOD,EMP}$ calculated with the parametric and the distance-based approaches. Each parametric model is able to well describe only a subset of the studied basins (see Figure 6), which is probably the reason why they demonstrate similar and non excellent performances when applied to the whole data set.

5. Conclusions

[39] The procedure for dimensionless flow duration curves estimation in ungauged basins developed in this work hinges on the concept of distance, that quantitatively represents the dissimilarity between curves and catchment's descriptors. This approach, based on distance matrices, allows one to account for a FDC as a whole object, avoiding the description of the curve by means of a parametric function. Moreover, no assumptions on the shape of the FDCs is made. This is an important feature when one has to manage at the same time curves described by a simple geometry (e.g., almost straight lines in the lognormal probability plot) and curves with more complex behavior (e.g., "S" shaped curves). In fact, complex shapes can be well described by a parametric model only using an high number of parameters, that sometimes cannot guarantee a robust parameters estimation.

[40] The results obtained by means of the distance-based model (nonparametric representation of the FDC) applied to our data set are comparable, and many times better, than the estimation yielded by classical parametric models of the same or greater complexity. These results are obtained on the basis of only two descriptors, while the lognormal model requires six descriptors for the assessment of two

Table 3. Descriptors Used to Estimate the Parametric Model's Parameters With Level of Significance and Variance Inflation Factor Test[a]

Model	Parameter	Descriptors	Student Test	VIF	R^2_{adj}
Lognormal	θ_1	asp, Cc	<0.05	<5	0.12
	θ_2	Xb, PLDP, slo, MHL	<0.05	<5	0.17
GPA	θ_1	Xmax, PLDP, slo, MHL	<0.02	<5	0.28
	θ_2	Ymin, IPS25, cos(or)	<0.05	<5	0.54
	θ_3	Xc, Yc, IPS50	<0.05	<5	0.39
PE3	θ_1	Xmax, PLDP, slo, Cc, MHL	<0.02	<5	0.39
	θ_2	Xc, Ymin, IPS100, Cc	<0.05	<5	0.31
	θ_3	Ymin, PS50	<0.02	<5	0.28

[a]Level of significance is measured by the Student test.

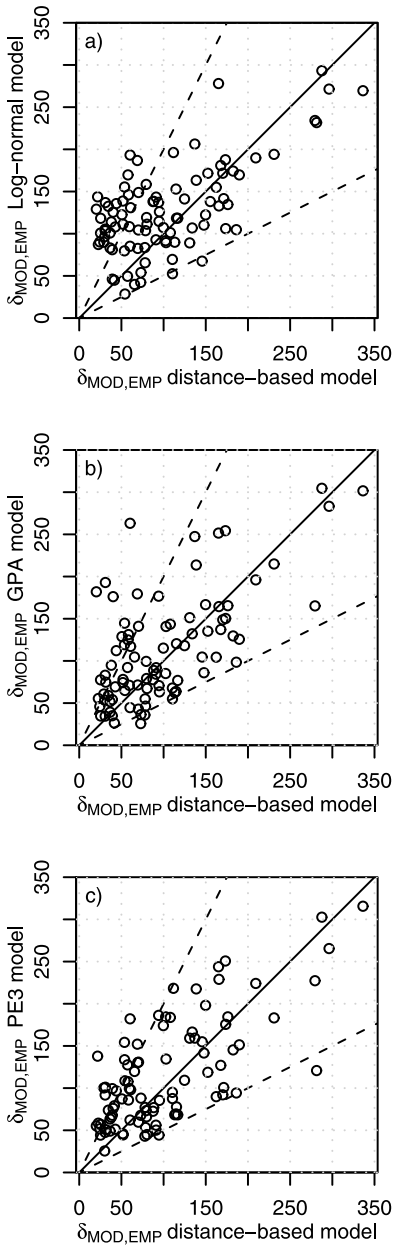

Figure 7. Quality of estimated dimensionless FDCs by the distance-based method compared with (a) the lognormal model, (b) the generalized Pareto model and (c) the Pearson type III model. The distance between the empirical curve and the estimated one, $\delta_{MOD,EMP}$, is reported in the scatterplot for each considered basin. The solid line represents the ratio 1:1 between the errors, while dashed lines delimit the areas where errors for the distance-based model are twice the parametric ones and vice versa. Points above the solid line represent curves better estimated by the distance-based method; points above the top dashed line represent curves much better estimated by the distance-based method.

parameters, and the PE3 and GPA models require 8 and 10 descriptors to estimate their three parameters.

[41] The main advantage of the method based on distance matrices is its ability in dealing with curves. For instance, the regionalization method proposed here could be improved considering also "complex" catchment descriptors as the hypsographic curve, or climatic information like the precipitation regime curve.

[42] **Acknowledgments.** The study has been supported by the Italian Ministry of Education through the grants 2006089189 and 2007HBTS85. Thanks are due to Martin Pfaundler from Federal Office for the Environment (BAFU), Switzerland, for provinding the data. Comments and suggestions from A. Castellarin, T. Torgersen, E. Martins, and two anonymous reviewers are gratefully acknowledged.

References

Burn, D. H. (1990), Evaluation of regional flood frequency analysis with a region of influence approach, *Water Resour. Res.*, *26*(10), 2257–2265.

Castellarin, A., G. Galeati, L. Brandimarte, A. Montanari, and A. Brath (2004a), Regional flow-duration curves: Reliability for ungauged basins, *Adv. Water Resour.*, *27*(10), 953–965.

Castellarin, A., R. M. Vogel, and A. Brath (2004b), A stochastic index flow model of flow duration curves, *Water Resour. Res.*, *40*, W03104, doi:10.1029/2003WR002524.

Castellarin, A., G. Camorani, and A. Brath (2007), Predicting annual and long-term flow-duration curves in ungauged basins, *Adv. Water Resour.*, *30*(4), 937–953.

Claps, P., and M. Fiorentino (1997), Probabilistic flow duration curvers for use in environmental planning and management, in *Integrated Approach to Environmental Data Management Systems*, *NATO ASI Ser., Partnership Subser. 2*, vol. 31, edited by N. B. Harmancioglu et al., pp. 255–266, Kluwer, Dordrecht, Netherlands.

CUBIST Team (2007), CUBIST project: Characterisation of ungauged basins by integrated use of hydrological techniques, *Geophys. Res. Abstr.*, *10*, 12048, sref:1607-7962/gra/EGU2008-A-12048.

Dalrymple, T. (1960), Flood frequency analyses, *U.S. Geol. Surv. Water Supply Pap., 1543-A*.

Fennessey, N., and R. Vogel (1990), Regional flow-duration curves for ungauged sites in Massachusetts, *J. Water Resour. Plann. Manage. Div. Am. Soc. Civ. Eng.*, *116*(4), 530–549.

Holmes, M., A. Young, A. Gustard, and R. Grew (2002), A region of influence approach to predicting flow duration curves within ungauged catchments, *Hydrol. Earth Syst. Sci.*, *6*(4), 721–731.

Hosking, J., and J. Wallis (1997), *Regional Frequency Analysis: An Approach Based on L-Moments*, Cambridge Univ. Press, New York.

Iacobellis, V. (2008), Probabilistic model for the estimation of *T* year flow duration curves, *Water Resour. Res.*, *44*, W02413, doi:10.1029/2006WR005400.

Kottegoda, N. T., and R. Rosso (1997), *Statistics, Probability, and Reliability for Civil and Environmental Engineers*, McGraw-Hill, New York.

Legendre, P. (1993), Spatial autocorrelation: Trouble or new paradigm?, *Ecology*, *74*(6), 1659–1673.

Legendre, P., F. Lapointe, and P. Casgrain (1994), Modeling brain evolution from behavior: A permutational regression approach, *Evolution*, *48*(5), 1487–1499.

Lichstein, J. (2007), Multiple regression on distance matrices: A multivariate spatial analysis tool, *Plant Ecol.*, *188*(2), 117–131.

Mantel, N., and R. Valand (1970), A technique of nonparametric multivariate analysis, *Biometrics*, *27*, 209–220.

Montgomery, D., E. Peck, and G. Vining (2001), *Introduction to Linear Regression Analysis*, 3rd ed., John Wiley, New York.

Mosley, M. P., and A. I. McKerchar (1993), *Handbook of Hydrology*, McGraw-Hill, New York.

NASA (2000), SRTM.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Found. for Stat. Comput., Vienna.

Rigon, R., and F. Zanotti (2002), *The Fluid Turtle Library, Users and Programmers Guide*, Univ. of Trento, Trento, Italy.

Singh, R., S. Mishra, and H. Chowdhary (2001), Regional flow-duration models for large number of ungauged Himalayan catchments for planning microhydro projects, *J. Hydrol. Eng.*, *6*(4), 310–316.

Sivapalan, M., et al. (2003), IAHS decade on predictions in ungauged basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences, *Hydrol. Sci. J.*, *48*(6), 857–880.

Smakhtin, V. (2001), Low flow hydrology: A review, *J. Hydrol.*, *240*(3–4), 147–186.

Smouse, P., J. Long, and R. Sokal (1986), Multiple-regression and correlation extensions of the Mantel test of matrix correspondence, *Syst. Zool.*, *35*(4), 627–632.

Viglione, A. (2007), nsRFA: Non-supervised regional frequency analysis, R package version 0.4-5, (Available at http://www.r-project.org/), R Found. for Stat. Comput., Vienna.

Viglione, A., P. Claps, and F. Laio (2007a), Mean annual runoff estimation in north-western Italy, in *Water Resources Assessment and Management Under Water Scarcity Scenarios*, edited by G. La Loggia et al., pp. 97–121, CSDU, Milan, Italy.

Viglione, A., F. Laio, and P. Claps (2007b), A comparison of homogeneity tests for regional frequency analysis, *Water Resour. Res.*, *43*, W03428, doi:10.1029/2006WR005095.

Vogel, R., and N. Fennessey (1994), Flow-duration curves. 2. New interpretation and confidence intervals, *J. Water Resour. Plann. Manage. Div. Am. Soc. Civ. Eng.*, *120*(4), 485–504.

Ward, J. (1963), Hierarchical grouping to optimize an objective function, *J. Am. Stat. Assoc.*, *58*, 236–244.

P. Claps, D. Ganora, and F. Laio, Dipartimento di Idraulica, Trasporti ed Infrastrutture Civili, Politecnico di Torino, I-10129 Torino, Italy. (pierluigi.claps@polito.it; daniele.ganora@polito.it; francesco.laio@polito.it)

A. Viglione, Institut für Wasserbau und Ingenieurhydrologie, Technische Universität, A-1040 Wien, Austria. (viglione@hydro.tuwien.ac.at)