Journal of Applied Water Engineering and Research, 2016
http://dx.doi.org/10.1080/23249676.2016.1196621

Taylor & Francis
Taylor & Francis Group

CASE REPORT

Flow duration curve regionalization with enhanced selection of donor basins

Muhammad Uzair Qamar[a*], Daniele Ganora[b], Pierluigi Claps[b], Muhammad Azmat[c], Muhammad Adnan Shahid[d] and Rao Arsalan Khushnood[e]

[a]*Faculty of Agricultural Engineering and Technology, Department of Irrigation and Drainage, University of Agriculture Faisalabad (U.A.F.), Faisalabad, Pakistan;* [b]*Department of Environment, Land and Infrastructure Engineering, Politecnico di Torino, Torino, Italy;* [c]*Institute of Geographical Information Systems (IGIS), School of Civil and Environmental Engineering (SCEE), National University of Sciences and Technology (NUST), Islamabad, Pakistan;* [d]*Water Management Research Center, University of Agriculture, Faisalabad, Pakistan;* [e]*NICE, National University of Sciences and Technology (NUST), Islamabad, Pakistan*

(*Received 3 March 2015; accepted 27 May 2016*)

A non-parametric regionalization procedure for the assessment of flow duration curve (FDC) at ungauged basins is presented. This modeling approach is fundamentally based on the quantification of dissimilarity between FDCs, thus allowing the grouping of most similar basins. An analogous grouping procedure, performed in the space of selected basin characteristics, allows the estimation of FDCs also at ungauged sites; however, for a fixed set of basin characteristics, some ungauged basins cannot be properly represented due to the scarcity of close (similar) donor basins. For these cases, the proposed method allows for the selection of an alternative set of basin characteristics as a support for similarity grouping. The results of the study show that the statistical error can be significantly reduced by following the proposed methodology. About 10% of all the basins involved in the analysis can benefit from the model swapping procedure, thus improving the final predicted curve.

Keywords: non-parametric; regionalization; flow duration curves; distance matrix; ungauged basin; distance-based method

1. Introduction

The data on flows in river, particularly of lower magnitude, are of great importance to meet the requirements of developmental projects for the management of water resources. It is also pertinent to mention that the problem of estimating hydrological variables in ungauged basins located in difficult terrain has remained the object of intense research activity in recent years (Razavi & Coulibaly 2013; Azmat et al. 2015, 2016). Different methods have been used to perform such estimation with the central idea of either extending or transferring the hydrological data, from gauged to ungauged sites, by complementing its relationship with catchment characteristics called descriptors (Botter et al. 2008; Blöschl et al. 2013). Among different statistics used to represent the low flow (see Smakhtin 2001 for a review), the flow duration curve (FDC) is a frequently used tool to represent the water availability at a river section; it is a cumulative-frequency curve which defines the relationship between magnitude of stream-flows of a certain time resolution (hourly, daily or monthly) and frequency of occurrence in any basin by translating the percentage of time for which a certain magnitude of flow equals or exceeds a certain flow value (Vogel & Fennessey1995).

Castellarin et al. (2013) have recently reviewed the currently used procedures aiming at the evaluation of FDCs in data-scarce basins, with a particular focus on the reliability of such methods in different climatic contexts.

The FDCs can be interpreted in two different ways (see, e.g. Vogel & Fennessey 1994; Castellarin et al. 2004a, 2004b, 2007): (i) total FDC (or period-of-record FDC), in which FDC of an individual station is constituted by the entire flow values occurring at that station and (ii) annual FDC, a separate FDC is constructed for individual years. Yokoo and Sivapalan (2011) disaggregated the FDCs into two components, i.e. slow FDCs and fast FDCs to develop a conceptual model to reconstruct FDCs, similar to the earlier work of Muneepeerakul et al. (2010) and Botter et al. (2007).

In practical applications, FDCs are often represented by a parametric function which is usually a probability distribution such as the generalized Pareto distribution with three parameters (Fennessey 1994), the Gumbel distribution (Kottegoda & Rosso 1997), the normal distribution (Singh et al. 2001) and the two- or three-parameter log-normal distribution (Fennessey & Vogel 1990; Claps & Fiorentino 1997), although in the past different non-probabilistic analytical forms have been popular (Müller et al. 2014). In more recent times, other distributions have also been used, for example, the Kappa (Castellarin et al. 2007), the EtaBeta (Iacobellis 2008) and the Burr type XII

*Corresponding author. Email: muhammad.uzair@uaf.edu.pk

(Ganora & Laio 2015). The choice of the distribution depends on the ability to adapt to the observed data and the possibility to estimate parameters in a robust manner. Parameters (or moments) of these probability distributions are evaluated in ungauged basin as a function of known basin characteristics (for instance, longitude/latitude of the basin centroid or morphologic, climatic and basin-scale features) (see Yusuf 2008). The hydrological model ModABa – MODel for Annual FDCs assessment in ephemeral small BAsins – has also been used recently for the assessment of FDCs by probabilistic characterization of the daily streamflow (Pumo et al. 2014, 2016). Another approach involves estimating the single FDC quantiles at fixed probability levels through independent regression models (Serinaldi 2011).

A different technique introduced by Ganora et al. (2009) grouped similar basins based on a dissimilarity index calculated between each pair of empirical FDCs. Prediction can be performed in a non-parametric way by averaging the dimensionless FDC of each element belonging to the cluster. However, this procedure can be non-optimal when there is no strong homogeneity within each group (Michele & Rosso 2002).

Whatever the model, regional analyses to estimate FDCs at ungauged sites are generally based on a few morpho-climatic characteristics which are not able to fully describe the complexity of the process, thus deteriorating the quality of the predictions (Laaha & Bloschl 2006).

In the present work, a possibility to improve the final estimate by adding a 'refinement' procedure to the first-try regionalized set is investigated. The analysis follows a procedure somewhat similar to that of Ganora et al. (2009) to obtain a first operational model for the whole case study in a strictly non-parametric way. Afterward, the space of the selected model is divided into an adequate number of clusters and models are re-selected for each cluster; a third step allows one to improve the modeling results by re-selecting a better model for points not well represented by the original cluster. This approach is, in particular, intended for those basins which are located away from the rest of the basins in the descriptor space; such basins will be referred to as remotely located basins (RLBs). For an RLB, any estimation made about their hydrological properties based on the neighbors can introduce an error into final calculations. To deal with these RLBs, a comparative analysis called model swapping procedure (MSP) is introduced and is discussed in detail in Section 3 after a general description of the distance-based procedure (Section 2).

The research work thus aims to investigate the following hypotheses: (i) whether dividing study area into smaller clusters and selecting separate model for each cluster can improve the results of predictions and (ii) whether the statistical predictions at RLBs can be improved by MSP.

2. Methodology

2.1. Data description

The time series dataset of 124 stations in Northwestern Italy (see Figure 1) has been considered for the application, having variable record length from a minimum of 5 to a maximum of 52 years, with a mean value of 12 years; the flow data are extracted from the publications of the former Italian Hydrographic Service extended with the more recent measurements provided by the Regional Environmental Agency (ARPA) of the Piemonte Region.

The whole database is described in Ganora et al. (2013) and is supported by an extensive collection of basin characteristics defined for each gauging station watershed (Gallo et al. 2013); such features include geomorphological characteristics obtained from the National Aeronautics and Space Administration Shuttle Radar Topography Mission (Farr et al. 2007) digital terrain model (pre-processed to a 100 m cell grid), climatic, vegetation and land-use descriptors (including soil characteristics) derived by properly clipping the thematic maps available for the area of interest. A summary of the range of some of the descriptors (out of 66 descriptors) used in the present research work is reported in Table 1.

Annual empirical FDCs have been constructed from the daily streamflow time series by ranking in the descending order the observations and associating to each value the exceedance frequency through the Weibull plotting position $F_i = i/(N+1)$, where i is the index of the sorted value and N the total number of observations in the year of interest. To allow an easier comparison of the curves in the dissimilarity-based framework, each yearly FDC has been resampled in order to have 365 values (see Figure 2 for sketching a curve with different number of elements resampled with a constant spacing along the frequency axis); this pre-processing does allow an easier handling of leap years and curves with missing values (in any case, only FDC with no more than 3 missing values per year has been used), while it does not affect the actual shape of the curve.

The final step in the data preparation is the computation of the average annual FDC, obtained by averaging the yearly curves at each frequency value F_i, thus obtaining a single FDC for each station (Vogel & Fenessey 1994; Mohor et al. 2015). Note that this curve represents a typical average year and its shape shows the typical within-year streamflow variability; a possible alternative is the period-of-record FDC, which accounts for all the available data and thus represents both the within- and the between-year variability. This research work focused only on the mean annual FDC as the time series have quite variable length (from 12 to 53 years) and records are not always overlapping. Moreover, for the effective application of the model, the data are made dimensionless by normalizing with the average flow value for that site as the focus is on the 'shape' of the FDC (the mean value is in general

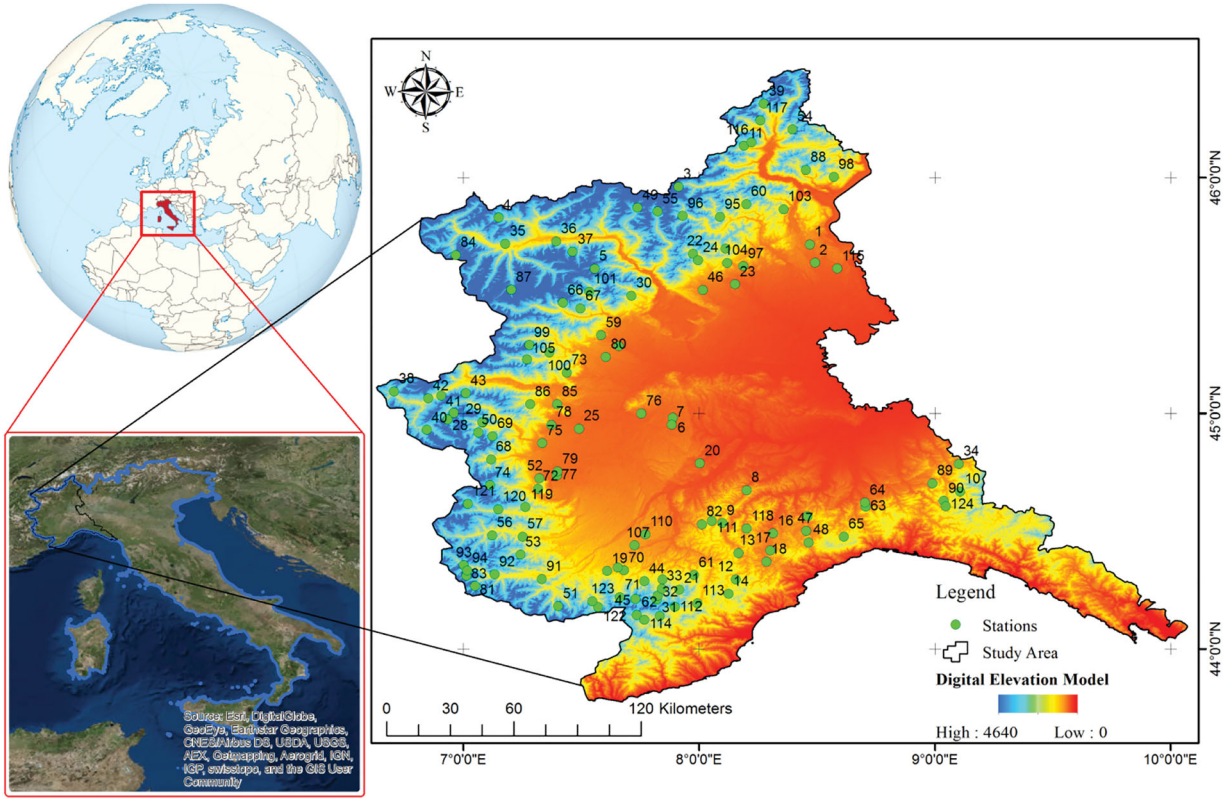

Figure 1. Location of gauging stations used in the analysis.

Table 1. Range of variation of some selected descriptors used in the distance-based procedure.

Descriptors (symbols)	Maximum	Mean	Minimum
Basin area (km^2) (A)	25,640	1276.33	22
Maximum elevation of the basin (m) (H_{max})	2750.55	4743	368
Latitude of basin (m) (Y_{cor})	5,129,050	4,977,667	4,886,350
Longitude of basin (m) (X_{cor})	508,450	401,454.8	319,450
Annual Normalized Difference Vegetation Index (NDVI)	0.644	0.451	0.082
Average basin elevation (m) (H_{mean})	2682	1306.51	244
Mean annual precepitation (mm) (MAP)	2183.037	1193.34	706.532
Percentage area of the basin as wetlands (clc_5)	7.89	0.181	0
Interquartile distance between basin elevation at 25% and 75% of area dominated by hypsographic curve ($delta_z$)	1762	695.346	50
Standard deviation of mean annual precepitation (mm) (MAP_std)	372.913	125.927	18.486
Percentage area of the basin which is not vegetated (clc_4)	78.68	16.038	0
Coefficient of variation in rainfall patterns (cv_{rp})	0.455	0.334	0.116
Time interval between maximum and minimum monthly averagesofrains($delta_{mon}$)	9	7.056	2
Percentage area of the basin as forests (clc_2)	90.56	48.54	6.2
Percentage area of the basin as herbaceous vegetation (clc_3)	89.32	33.15	8.57
75th percentile of the hypsographic curve (a_{75})	2462	941.76	202
Coefficient of rainfall intensity (C_{int})	0.036	0.0199	0.011
Standard deviation of exponent of intensity - duration - frequency curve (IDF_{a-std})	37.88	23.39	11.88
Average values of coefficients of the Fourier series representing rainfall patterns ($fourier_{B1}$)	49.56	− 7.94	− 56.55

easier to estimate, by using a few years of observations or simple models) (Smakhtin & Weragala 2005; Poff et al. 2006).

2.2. *Selection of relevant descriptors*

The first step of the regionalization procedure is analogous to that adopted by Ganora et al. (2009); a dissimilarity

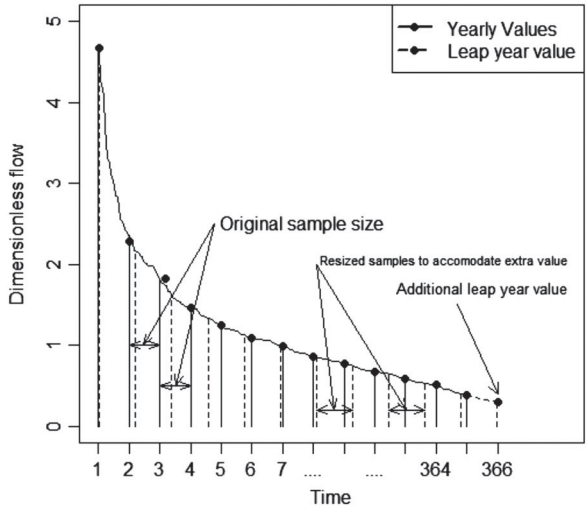

Figure 2. Sampling points with original and increased sample sizes to accommodate leap year values.

matrix is used to store the dissimilarity measure between all the pairs of FDCs. The dissimilarity between FDCs of any two stations is calculated by the use of the following simple equation:

$$D_a = \sum_{i=1}^{365} |Q_{i,s} - Q_{i,r}|, \quad (1)$$

where the value D_a is total magnitude of dissimilarity between FDCs of stations 's' and 'r' having flow magnitude of $Q_{i,s}$ and $Q_{i,r}$, respectively, being i the index of the FDC value.

The descriptors are normalized by using the mean values of each descriptor for the entire study area (see Reed et al. 1999; Livneh & Lettenmaier 2013). Afterward, the distance matrices of descriptors are obtained by calculating dissimilarity (D_d) between descriptors values (d_s and d_r) at two sites; if the descriptor is a single-value number (e.g. basin area, mean elevation), the dissimilarity is simply the absolute difference of the descriptors values of the two basins, i.e. $D_d = |d_s - d_r|$; otherwise, if the descriptor is a curve itself (e.g. hypsometric curve), Equation (1) can be used (Qamar et al. 2015). The approach has the advantage of enabling the use of the whole curve as a single variable with respect to represent the curve through fixed quantiles or parameters of analytical representation. The FDC distance matrix can be correlated, by means of linear regression models, to the distance matrices of one or more descriptors and properly tested to find the significant correlations. The complete information regarding the regionalization procedure is presented in the form of flow chart in Figure 3.

Since the aim is to convert the descriptors' data into hydrological data, therefore the representative descriptors (dominating descriptors) should be first defined. To start with, the distance matrices for each descriptor D_{d_i} as well as for the FDCs D_Q are firstly determined.

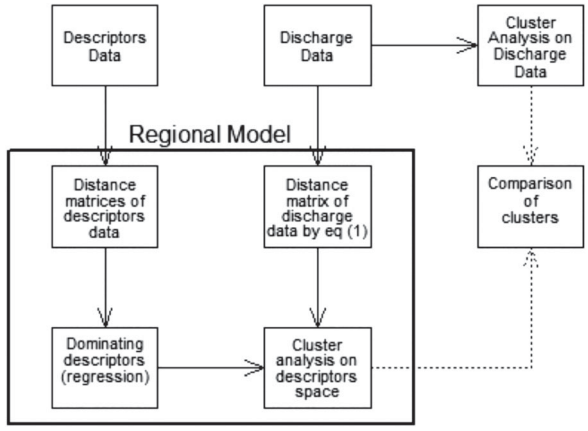

Figure 3. Schematic diagram representing the steps involved in the regionalization procedure followed by Ganora et al. (2009).

The dominating descriptors are bracketed by their relationship with FDCs. The multiregressive approach is used to assess the relationship between distance matrix of discharge and descriptors; the statistical model can be written as

$$D_Q = \beta_1 D_{d_1} + \beta_2 D_{d_2} + \beta_3 D_{d_3} + \cdots + \beta_p D_{d_p} + C_0, \quad (2)$$

where D_Q and D_d are distance matrices of discharge and descriptors, respectively, unfolded to be represented as vectors (Lichstein 2007); p is the number of descriptors involved; Δ are the regression coefficient; and C_0 is residual matrix. The strength of regression is determined by

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p-1}, \quad (3)$$

with n as number of basins and R as the standard coefficient of determination (e.g. Kottegoda & Rosso 1997).

Due to the large number of regressors, it is lively to find models with a non-negligible correlation between descriptors. In these cases, the variance inflation factor (VIF) which measures the undesirability of multicollinearity in a least-square regression analysis becomes unignorable. It quantifies, through an index estimation, the inflation occurred in variance of an estimated regression coefficient. A cut-off value of 5 is used, beyond which the selected model is dropped (Montgomery et al. 2001).

Analogously to Ganora et al. (2009), the present methodology used the procedure defined by Lichstein (2007), which provides a method for multiple regression on distance matrices (MRM) and an extension of the Mantel test to check the significance of regression coefficients when working with distance matrices (analogous to the t-Student test for regular regressions).

Regression models that passed the Mantel test and VIF test, and having the higher R^2_{adj} values, are used to implement a cluster analysis procedure in order to provide adequately homogeneous (in statistical sense) pooling groups.

Table 2. Descriptor models for the overall study area and clusters enlisted in the order of Δ values (models in *italic font* are the selected ones).

Model vicinity	Descriptors (symbols)	R^2_{adj}	VIF	Delta factor
Overall	clc_5, H_{max}	0.024	< 5	59.00
	y_{cor}, hypsographic curve	*0.041*	*< 5*	*59.66*
	H_{max}, NDVI	0.030	< 5	60.26
	NDVI, $delta_z$	0.023	< 5	60.48
	y_{cor}, H_{mean}	0.033	< 5	60.82
Cluster-1	MAP_std, NDVI	*0.035*	*< 5*	*57.33*
	clc_4, cv_{rp}	0.048	< 5	58.85
	H_{max}, NDVI	0.038	< 5	59.06
	clc_4, clc_5	0.047	< 5	59.74
	H_{max}, hypsographiccurve	0.049	< 5	59.84
Cluster-2	$H_{max}, delta_{mon}$	0.066	< 5	51.46
	clc_5, H_{max}	0.069	< 5	53.66
	H_{max}, NDVI	*0.065*	*< 5*	*56.17*
	MAP, C_{int}	0.029	< 5	56.24
	H_{max}, C_{int}	*0.066*	*< 5*	*57.64*

For each selected model (within each cluster), each station is considered, in turn, as ungauged and its FDC is estimated by the nearest neighbors (NNs) (please refer to Section 2.3) and then compared to the original empirical one. The error measure ζ, defined as the dissimilarity (computed by Equation (1)) between the empirical and the estimated FDCs, is calculated for each station. The overall error measure $\Delta = \sum_1^n \zeta$ is then used as a comprehensive performances index of the specific model and used to rank the different regressions. The best models are the combinations of descriptors which generate the lowest values of Δ and comparatively good R^2_{adj} values. The R^2_{adj} values obtained with regression models with distance matrices are low, although the descriptors result to be statistically significant. Lower R^2_{adj} values arise from simpler models with only two descriptors, as in Table 2.

The proposed methodology of distance-based measurement is carried out in the R statistical environment (R Development Core Team 2013), desegregated for Mantel test and multivariate regression analysis in the nsRFA package (Viglione 2007).

2.3. Implementation of the model

Based on the descriptors' distance matrices selected above, the empirical FDCs of neighboring basins are used to execute the FDCs for ungauged basin. There are different procedures available in the literature to choose the neighboring basins, for example formation of fixed regions through cluster analysis (Hosking & Wallis 1997; Viglione et al. 2007) or based on region of influence (ROI) (Burn 1990). The present methodology use a combination of classification techniques grounded on the dissimilarity-based approach depicted above: cluster analysis, performed on the basis of dominating descriptors selected in the first step, is followed by ROI grouping in each cluster to assess FDCs in ungauged basin. The Ward agglomerative hierarchical algorithm (Ward 1963) is used as it is able to generate compact clusters with evenly distributed basins in each of them. In this work, the stations are clustered on the basis of dominating descriptors and each cluster is treated as a separate entity (no reallocation or homogeneity test for cluster independence is considered). The number of clusters (two macro sub-regions) is selected by using NbClust package in the R statistical environment, which provides best clustering scheme by observing the results obtained by varying the number of clusters, distance measurement and clustering technique (Charrad et al. 2014). The aim of doing cluster analysis is to select a model for each region executed by dividing the entire study area purely on the descriptors values to reduce the error magnitude resulting from the extension of single model over the whole study area (Laaha & Bloschl 2006).

However, the definition of the optimal number of clusters is a difficult task because one looks for a few homogeneous groups with a large number of elements; these conditions are rarely obtained as homogeneity tends to increase with decreasing number of elements within the cluster (and thus with increasing the number of clusters). The rationale of this work is to accept clusters even if they are slightly non-optimal (thus avoiding complicated procedures such as reallocation of elements between clusters) and move the optimization step to the second phase of the procedure.

The methodology is applied to the basins with areas varying from 22 to 25,640 km^2. For the sake of simplicity in interpretation of results, the basins are classified on the basis of their areas. The basins with an area less than 500 km^2 are considered small; medium-sized basins have their areas ranging between 500 and 1000 km^2 and those which surpass 1000 km^2 are classified as large basins.

The results generated by the proposed model for the entire study area are tested by using a cross-validation procedure. It is done by considering one station as ungauged, removing it from the whole database and estimating FDC for that station with the proposed approach. For prediction, the models with two descriptors are preferred because of their higher robustness and for an ease of comparison with geographical distance method (explained in Section 4.2), which is also a combination of two descriptors (latitude and longitude). The process is repeated for all the stations and the errors are measured between estimated FDCs and empirical FDCs.

3. Model swapping

Whatever the method adopted to group basins, in any descriptor space constituted by the selected variables, the stations located away from the rest of the basins may not be well represented in terms of neighbors; these are the RLBs, and an example is reported in Figure 4 (red dot). When

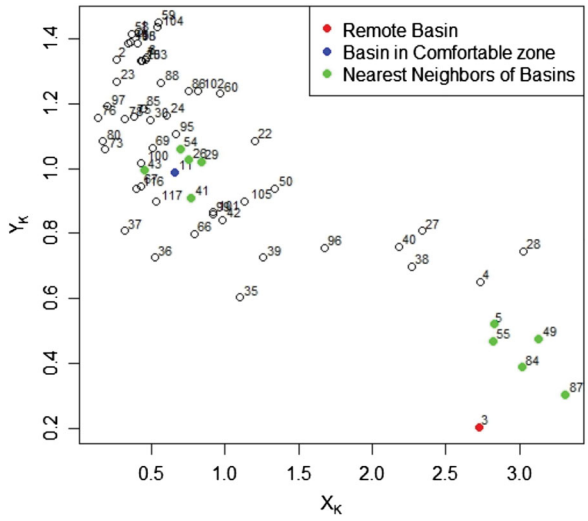

Figure 4. Station represented by red dot is located remotely ($S^3_{\mathrm{NN}}/\mathrm{TD}_n = 2.849 > 1.50$) and the one in green is located in comfortable zone ($S^{11}_{\mathrm{NN}}/\mathrm{TD}_n = 0.68 < 1.50$), in space of any selected model (x_K, y_K).

hydrological data are assessed at RLBs, there is always a considerable amount of error in the final calculation due to scarcity of data, which prevents the usage of standard models (Pellicciotti et al. 2012). In the space of dominating descriptors, there can be stations having rather different descriptors values from the rest of the sample and since the values of descriptors directly affect the hydrological properties of basins, therefore the assessment made on their hydrological properties from its neighboring basins can introduce a reasonable amount of error (Pechlivanidis et al. 2014). Hence, it can be concluded that the selected model is good for the majority of basins, but it is likely to be not representative of the remote ones.

In hydrological modeling, the relationship between descriptors and hydrological data is generally considered through R^2_{adj}, which represents the quality of fit for the model (e.g. Heerdegen & Reich 1974; Mimikou & Gordios 1989; Post & Jakeman 1996, 1999; Tung et al. 1997; Sefton & Howarth 1998; Pandey & Nguyen 1999). This selection criterion is backed up with the Δ factor, which indicates the overall error magnitude produced by the model. Based on the model selection criteria defined in section 2.2 (R^2_{adj} and Δ), the hypothesis is that for any RLB (say X) in the selected model space having certain R^2_{adj} and Δ factor values can be replaced by another model with similar performances, but where X is in a more desirable position among its neighbors (e.g. arranged approximately in the middle of the cluster of stations), called *comfortable zone* (see blue dot in Figure 4).

The definition of comfortable zone comes from Korn and Muthukrishnan (2000), who are the first to study reverse k NNs (R kNNs) queries. The R kNN of any query point (say P) executes the objects in the database which have P as their NN. Later, Stanoi et al. (2000) solved the R kNN queries by partitioning the whole space around the data point into six equal regions (each of 60°). The same concept is used in defining the confidence zone of an unknown data station in the present work with a slight modification, i.e. instead of finding all the stations that have unknown data station (P) as their NN, the NNs of P are found because the former can result in only a limited or no NN (in case of an RLB) leading to oversimplified or no result, respectively.

The MSP assumes that each NN contributes equally to the hydrological data of ungauged basin. The hydrological dissimilarity between two basins is proportional to the distance between them in the selected descriptors model space (Ganora et al. 2009). Therefore, theoretically, the best location for an ungauged basin is to be in the middle of its neighbors because only then, each station is at an equal distance from the unauged basin and is contributing equally. The efficient estimations are acquired if the hydrological average of the NNs of ungauged basin on descriptors space, which hypothetically locates ungauged basin at the centroid of NNs, is also matched by the location of ungauged basin in descriptors space among its neighbors.

The MSP developed here is a comparative analysis and is valid only for the stations which are remotely located. No mathematical definition of RLB is present in the literature for this kind of application. In the present work, the following procedure is used to define an RLB:

(1) A comparison of station-neighbors distance for any selected station, say X, with station-neighbors distances of rest of the stations.

If S^X_{NN} is the sum of station-neighbors distances for the basin X, then for more general case of n number of basins, it can be written

$$\mathrm{TD}_n = \frac{\sum_{i=1}^{n}(S^i_{\mathrm{NN}})}{n}, \tag{4}$$

where TD_n is average station-neighbors distance for the entire basins in the study area. The basin X can be reasonably considered an RLB if $S^X_{\mathrm{NN}} > 1.5 \cdot \mathrm{TD}_n$.

The multiplication factor of 1.5 is used to relate S_{NN} and TD_n. The threshold is carefully selected among a number of available options (e.g. 1.10, 1.25, 1.50, 2.00, and 3). The lower threshold (e.g. 1.10) substantially increased the number of RLBs, thus undermining the importance of initially selected model. Moreover, RLBs should have a clear dissimilarity magnitude in terms of descriptors values from remaining basins in the study area. For lower threshold, this pattern may not be achieved due to smaller difference between the values of S_{NN} and TD_n. On the contrary, the higher threshold (e.g. 2 or 3) executes only a limited or at times no RLB.

(2) Observing the neighbors in six sectors around the station (Stanoi et al. 2000).

Generally, due to the unique position of remote basin in the study area, its NNs are either concentrated in one of the six regions around it or basin is not covered from all sides (see Figure 4). The swapped model should increase the covering of basin by its neighbors.

Practically speaking, referring to Figure 4, the orientation of NNs for station # 11 (blue filled point) is more desirable than that of station # 3 (red filled dot). Therefore, the statistical estimates of station # 3 can be improved by MSP.

It should be noted that the selection of model to bring an RLB in a comfortable zone is done by swapping it with a model having similar Δ and R^2_{adj} values, on the exact same procedure as discussed in previous section for the selection of model for each cluster.

After performing statistical tests, the models with two descriptors for clusters and overall study area are selected. Ideally, each descriptor value of each station should be uniformly scattered over the entire space of a selected model. The present methodology employed the use of density plots to measure the 'degree of scatter' of each descriptor values.

The difference between Δ values of the original and the swapped models should be as lower as possible (not greater than 5% in the present methodology) in order to consider the two models showing similar performances; higher the difference between the delta values, higher will be the error in the swapped model space. The higher error in the swapped model will obviously make its prediction more unreliable even with the increased coverage.

To clarify the concept of MSP to have better spatial coverage around the ungauged data point, an example is reported. The statistical results of the NN analysis, for station # 4 (represented with red filled dot in Figure 5) and station # 45 (represented with red filled dot in Figure 6), before and after improving the spatial coverage of neighbors (represented with green filled dot) are compared. The selected five models for overall study area and selected clusters are enlisted in Table 2, which carries Δ values and R^2_{adj} of the models. The outputs of originally selected and swapped models in terms of root mean square error (RMSE), Nash–Sutcliffe efficiency (NSE) and mean absolute error for the considered station are represented in Table 3.

To summarize, the complete procedure for the evaluation of FDCs requires: (1) the computation of the distance matrix of the FDCs; (2) the computation of the distance matrix of each descriptor; (3) the Δ and MRM will give us operational models for the overall study area, (4) the space of the selected model for the overall study area is divided into smaller cluster, (5) based on the previously defined procedure of model selection, model is selected for each smaller region, (6) RLBs are shifted to comfortable zone by utilizing MSP, and (7) the regional dimensionless FDCs are estimated by NN method.

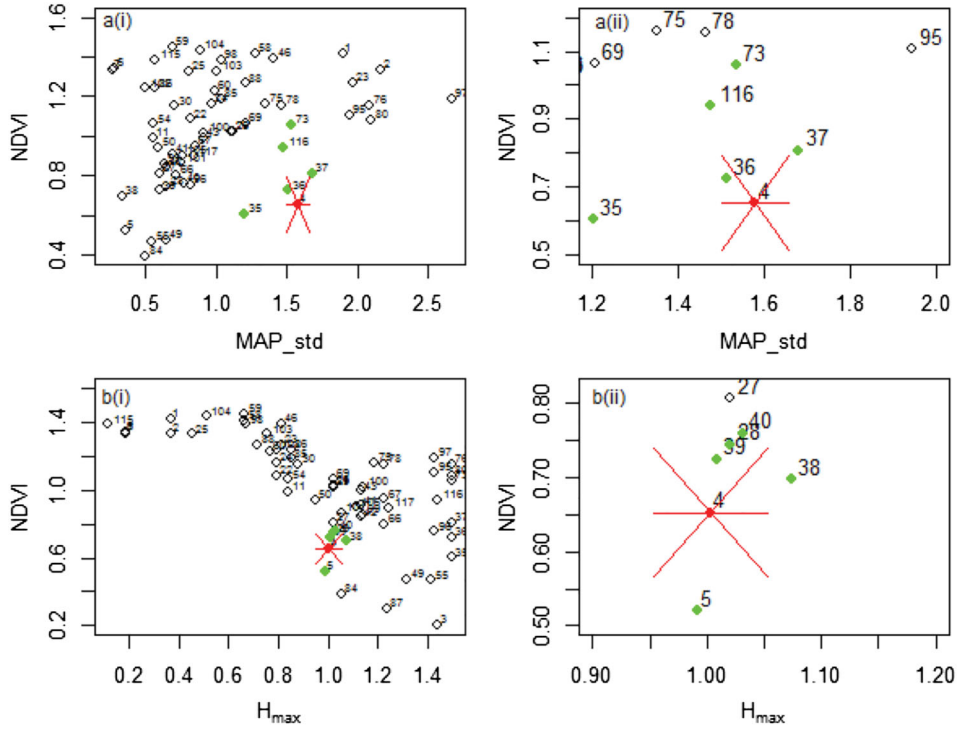

Figure 5. (a(i)) Basins arrangement in the space of selected model; (a(ii)) detailed view of selected basin and its neighbors; (b(i) and b(ii)) swapping model to give better neighbor coverage and its detailed view, respectively.

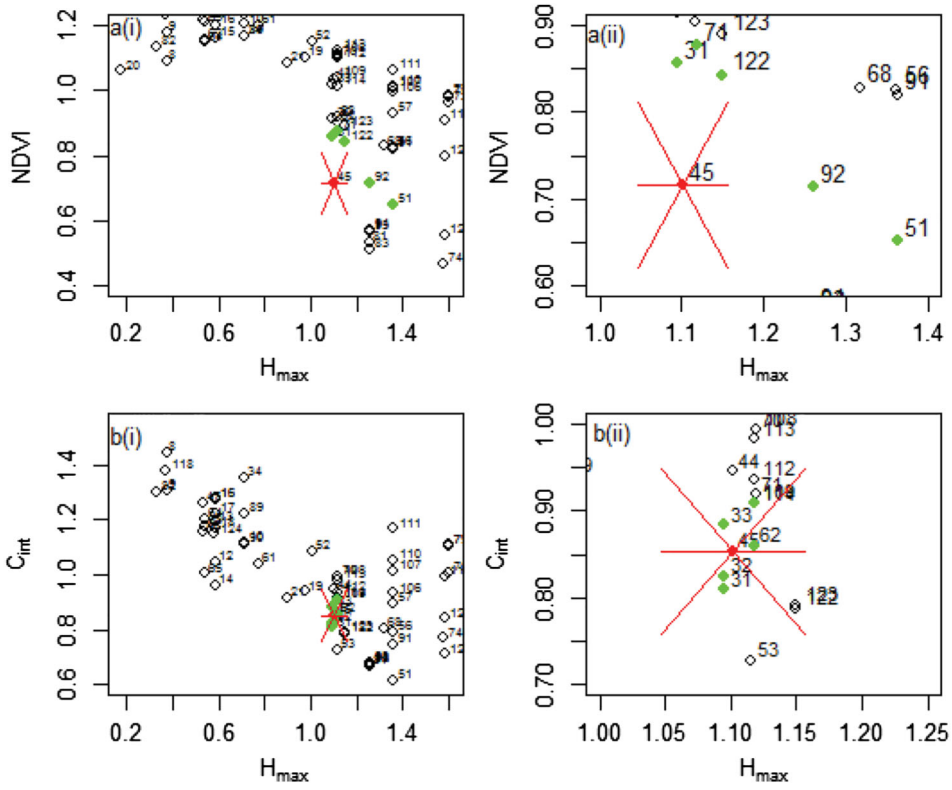

Figure 6. (a(i)) Basins arrangement in the space of selected model; (a(ii)) detailed view of selected basin and its neighbors; (b(i) and b(ii)) swapping model to give better neighbor coverage and its detailed view, respectively.

Table 3. Results generated by original and swapped model in terms of RMSE, NSE and D_a.

Basin number	Cluster number	Original model	RMSE	NSE	D_a	Swapped model	RMSE	NSE	D_a
4	1	MAP_std, NDVI	0.144	0.960	34.071	H_{max}, NDVI	0.075	0.989	21.492
45	2	H_{max}, NDVI	0.214	0.949	59.600	H_{max}, C_{int}	0.155	0.973	41.972

4. Alternative methods

The proposed procedure has been applied to the Italian case study and compared to the results produced by other regional approaches, to verify the suitability of the results. The alternative methods used in the comparison are a parametric model and a non-parametric approach based only on the geographical proximity of the donor stations.

4.1. Parametric model

For comparison, the results obtained from a regional model, based on the Burr distribution, developed (fitted to observed data and predicted at ungauged stations) over the same dataset of this study (Ganora et al. 2013) are used. The parametric model provides the FDC at ungauged sites according to the quantile equation of the Burr, which reads

$$x(P) = a\left(\frac{(1-P)^{-b}-1}{b}\right)^{1/c}, \qquad (5)$$

where a is the location parameter, while b and c are the two shape parameters. For particular conditions, the Burr becomes a two-parameter Weibull ($b \to 0$), or a two-parameter Pareto ($b \to -\infty$ and $c \to \infty$ jointly).

Fitting of the model is performed by first computing sample L-moments for all the available mean annual FDCs. L-moments are widely used statistics that, analogously to ordinary moments, contain information about the average, the variability, the skewness, etc., of a distribution. For further details about theory and application of L-moments, the reader is referred to Hosking and Wallis (1997) and references therein. To evaluate the set of L-moments in ungauged basins, Ganora and Laio (2015) used multiple regressions based on catchment descriptors adequately selected among a wide database of basin characteristics. The choice of the 'best' models is performed by checking all the possible regression models with a combination of 1–4 descriptors: first, the models are tested for significance (t-Student test) and multicollinearity (VIF test); then, the models are sorted according to their prediction

performances. Final model reads:

$$Y = -7.3605 \cdot 10^{2} + 1.2527 \cdot \mathrm{MAP} + 3.2569 \cdot 10^{-1} \cdot H_{\mathrm{mean}} + 5.2674 \cdot \mathrm{fourier}_{B1} - 6.7185 \cdot \mathrm{clc}_{2}, \quad (6)$$

$$\mathrm{LCV} = -2.896 \cdot 10^{-1} - 2.688 \cdot 10^{-3} \cdot \mathrm{clc}_{3} + 9.643 \cdot 10^{-5} \cdot a75 + 1.688 \cdot 10^{-4} \cdot \mathrm{MAP} + 2.941 \cdot 10 \cdot c_{\mathrm{int}}, \quad (7)$$

$$\mathrm{LCA} = 4.755 \cdot H_{\mathrm{max}}^{-0.2702} \cdot \mathrm{IDF}_{a\_\mathrm{std}}^{0.06869} \cdot \mathrm{cv}_{\mathrm{rp}}^{0.2106}, \quad (8)$$

where Y is the mean annual runoff in mm, L-CV is the L-coefficient of variation and L-CS is the L-coefficient of skewness. The descriptors used in Equations (6)–(8) are defined in Table 1, while details about descriptors can be found in the original publication (Ganora & Laio 2015). However, it is recalled that H_{mean}, H_{max} and $a75$ are morphological indexes; MAP, $\mathrm{fourier}_{B1}$, C_{int}, $\mathrm{IDF}_{a-\mathrm{std}}$ and $\mathrm{cv}_{\mathrm{rp}}$ refer to climatic features at the basin scale; and clc_2 and clc_3 are land-use characteristics.

The FDC prediction in ungauged basins is finally performed by computing the L-moments on the basis of the descriptors of the target basin through Equations (6)–(8), and thus calculating the parameters a, b and c of Equation (5) from the L-moments. Relationships between parameters and L-moments are not reported for brevity, but numerical methods as well as approximate formulas can be found in Ganora and Laio (2015).

This regional approach can be interpreted, following the classification of Wagener et al. (2007), as a 'mapping function' as it does not require homogeneous regions but only an interpolating function (the regression in this case) that is assumed valid over the whole case study areas (see also Laio et al. 2011 for an application of the same idea to flood frequency analysis).

4.2. *Geographical distance method*

An alternative non-parametric method to regionalize the FDC is the selection of NNs according to the geographical distance. It is assumed that the stations having similar hydrological properties are located closer to each other in the geographical space and hence it is reasonable to asses hydrological properties of ungauged catchments based on spatial proximity (Blöschl 2005). The Euclidean distance norm is generally used to calculate distance between a pair of catchment centroids and is adopted for this study.

A preliminary processing (not shown) has reported that average of the nearest (in geographical space) five neighbors provides better prediction performances for the case study.

5. Result

Before presenting the results of the study, it is worth commenting that only a few of the many parameters available to describe the basin morphology and climate are required by distance-based method for efficient estimation of FDC in an ungauged basin. Figure 7 reports the actual and predicted FDCs in cross-validation. It can be observed from Figure 7 that the agreement between actual and predicted FDCs is well established in case of the proposed method when compared to the other methods.

As performance indexes, RMSE, D_a and NSE are evaluated. These performance indexes for the four considered procedures (proposed model, geographical distance method, parametric model and by the usage of single model for entire study area) are listed in Table 4 for some randomly selected basins, while a complete comparison is shown in Figure 8. The solid line, in Figure 8, represents the ratio 1:1 between the errors, while dashed lines delimit the areas where errors for the proposed method are twice the parametric ones and vice versa. Points above the solid line represent FDCs better estimated by the distance-based method; points above the top dashed line represent FDCs much better estimated by the proposed method.

Since large number of basins in Figure 8 are located above the solid black line, it can therefore be interpreted that the proposed methodology works better than the other comparative methods.

The newly developed method, in which models are predominately constituted by climatic descriptors due to better global performance, out-performed the other methods in medium- and large-sized basins but in some of the small basins the error magnitude is comparatively large. The results are in line with the major finding of the Land-Ocean Interaction in Coastal Zone research which states that due to modulation capacity of larger rivers the influence of climatic descriptors affects the smaller basins more dramatically than larger basins (Crossland et al. 2005). Generally, the proposed model performs well for the majority of basins in the study area.

The latter part of the work involves the identification of the basins which are located away from the rest of the basins. By using set criteria (Equation (4) and NNs in six sectors), some of these basins are identified as RLBs. It is for these basins that MSP is introduced. Of the entire basins used in analysis; 10% basins are found to be remotely located whose prediction is improved using the MSP (see Figure 9).

Figure 9 shows that the statistical error of estimation is significantly reduced by swapping the initially selected model with another model, based on set criteria.

The methodology is served well by the models constituted by the climatic-geomorphological characteristics, while certain land-use characteristics (e.g. clc_5) and extended information from rainfall pattern (e.g. $\mathrm{delta}_{\mathrm{mon}}$) produced, comparatively, less appreciable results. The

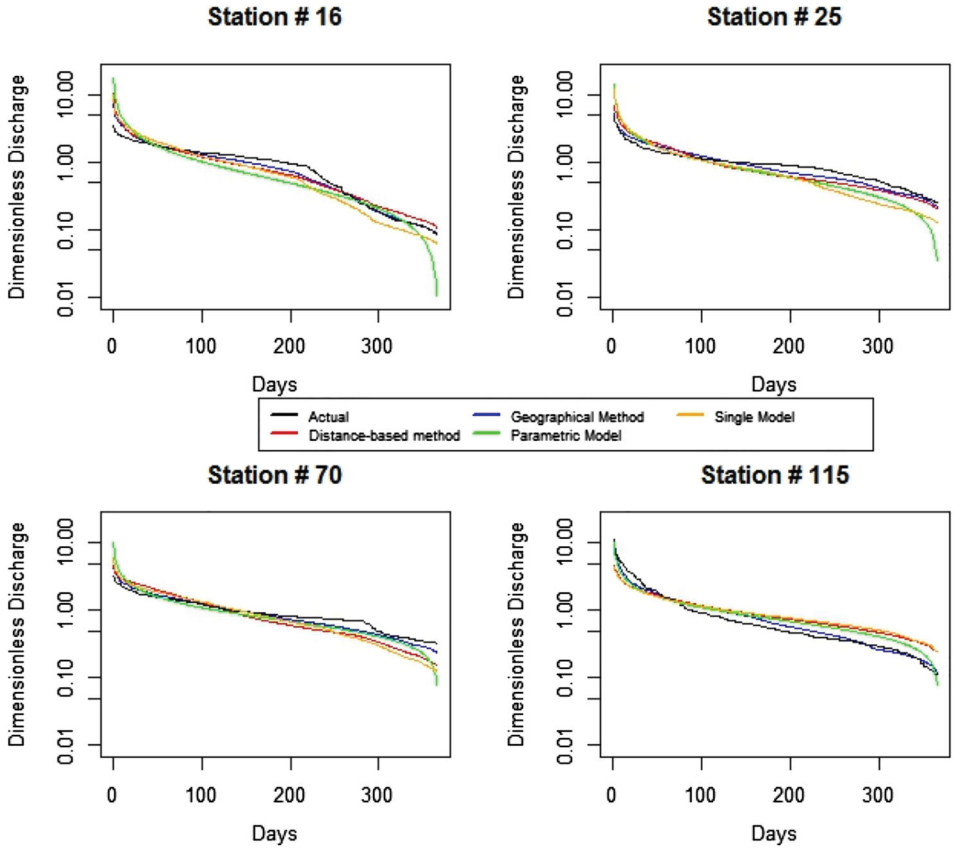

Figure 7. Comparison of simulated FDCs with actual FDCs at selected stations.

Table 4. RMSE, NSE and D_a obtained in the various basins using different methods.

Basin Station codes	Area (km^2) A	New method			Euclidean (geographical) distance method			Parametric (Burr) model			Single model for entire study area		
		RMSE	NSE	D_a	RMSE	NSE	D_a	RMSE	NSE	D_a	RMSE	NSE	D_a
1	204	0.050	0.995	13.370	0.077	0.098	23.401	0.734	0.087	78.307	0.121	0.975	38.135
2	382	0.190	0.862	42.825	0.297	0.667	73.827	0.950	− 2.40	128.99	0.264	0.735	63.378
4	69	0.075	0.989	21.492	0.182	0.936	56.814	0.3116	0.815	63.079	0.204	0.921	65.797
7	350	0.240	0.965	33.921	0.459	0.873	79.694	0.884	0.530	80.096	0.254	0.961	43.354
13	496	0.101	0.996	12.690	0.066	0.996	12.700	0.445	0.821	64.498	0.132	0.984	22.845
15	2594	0.119	0.983	27.609	0.257	0.921	44.320	0.874	0.089	93.425	0.251	0.924	61.291
27	154	0.123	0.975	23.373	0.287	0.866	40.990	0.428	0.703	59.176	0.120	0.966	26.682
48	91	0.209	0.910	61.102	0.499	0.491	84.939	1.146	− 1.682	190.120	0.735	− 0.103	122.20
60	147	0.162	0.938	34.744	0.164	0.936	36.585	0.278	0.933	56.426	0.370	0.678	73.693
69	953	0.268	0.944	65.428	0.257	0.948	69.384	0.455	0.840	109.697	0.364	0.898	94.980
76	25,640	0.161	0.886	40.613	0.486	− 0.029	93.630	0.297	0.877	41.215	0.329	0.526	82.503
88	122	0.212	0.956	45.702	0.316	0.903	84.129	0.659	0.580	66.379	0.274	0.927	80.897
92	560	0.190	0.901	40.000	0.192	0.902	41.963	0.439	0.485	50.929	0.228	0.861	69.134
105	233	0.117	0.983	34.424	0.210	0.948	65.967	0.456	0.757	70.506	0.251	0.926	72.625
109	249	0.051	0.994	9.763	0.130	0.966	37.304	0.501	0.504	78.204	0.153	0.954	43.982
112	375	0.182	0.900	38.240	0.274	0.773	67.418	0.365	0.782	63.615	0.272	0.778	59.710

apparent reason seems to be the difficulty in nominating the unique NNs of ungauged basin due to the lower degree of scatter of the descriptor values in the space of selected model (i.e. many basins with exact same descriptor values). Figure 10 reports the density plots of two descriptors NDVI and delta$_{mon}$ (left and right, respectively). The plot of NDVI is uniformly scattered with each basin having a unique descriptor value, whereas the plot of delta$_{mon}$ shows that the stations are clustered at a particular section of the descriptor space and a number of stations have similar

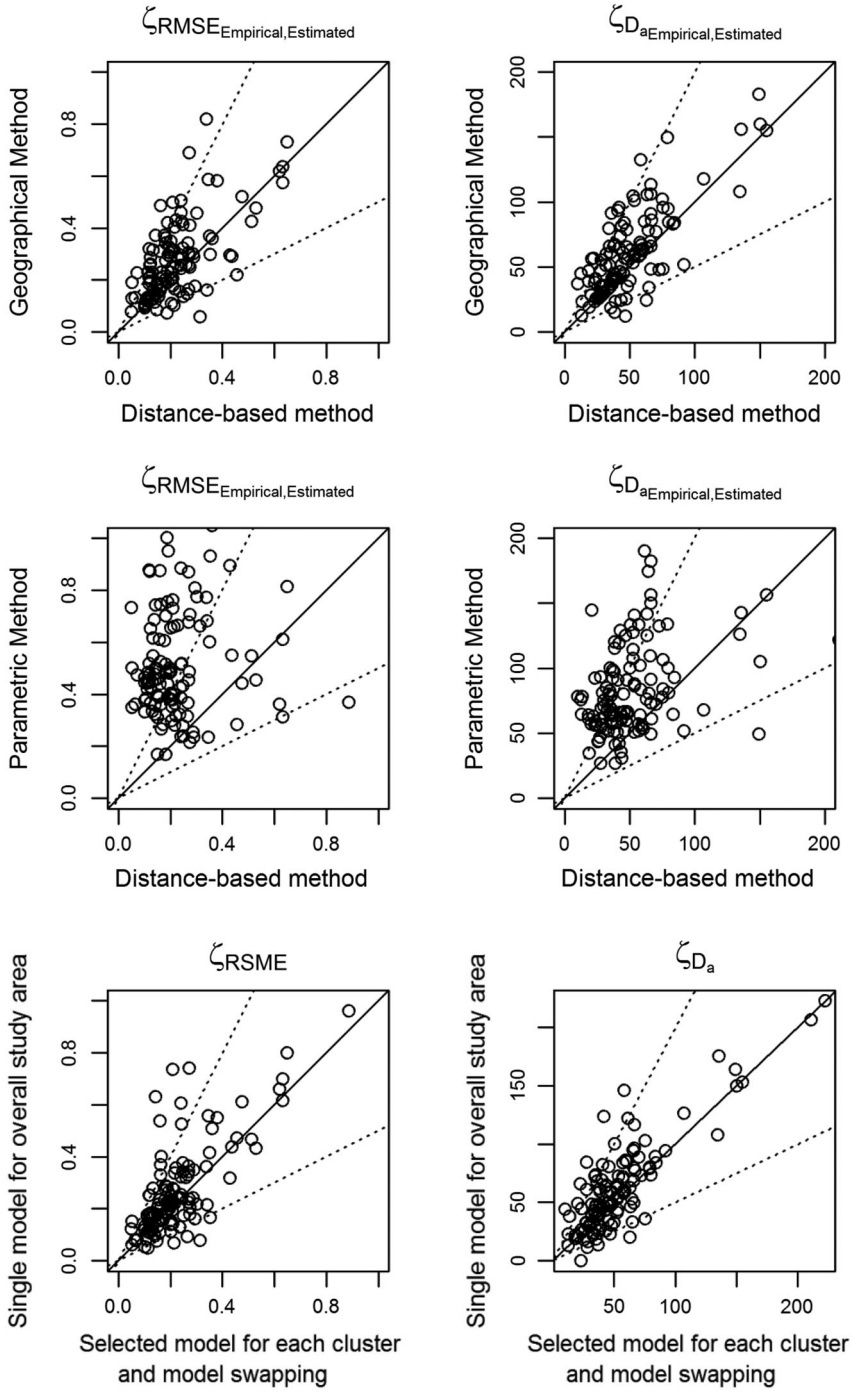

Figure 8. Comparison of RMSE and D_a of distance-based model, geographical method, parametric model (Burr distribution) and single model for overall study area.

descriptor values, making it difficult to differentiate the basins. Therefore while applying this method, it is advisable to use those basin characteristics, which allow easy identification of unique NNs of ungauged basin.

The size of the Δ factor value is constrained in MSP (i.e. the value should not exceed 5% of the initial model selected for the cluster) because larger Δ value differences exaggerate the error magnitude in the space of selected model, making the executed results for any station unreliable.

The comparison in Table 5 shows that the performance of a model is considerably enhanced (in terms of RMSE, NSE and D_a) by dividing the study area into clusters and selecting individual model for each of it.

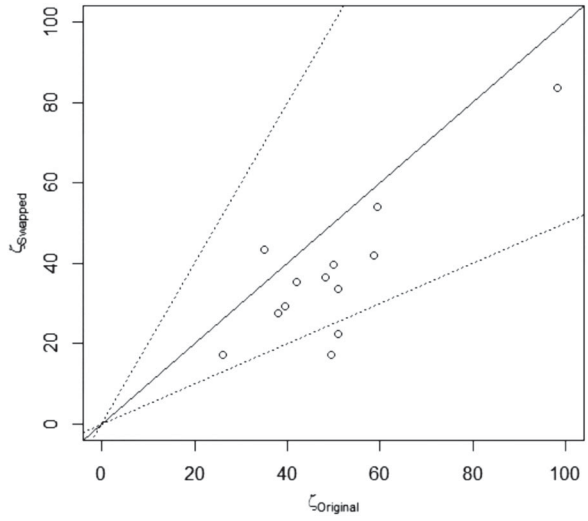

Figure 9. Comparison of error magnitude between originally used model and swapped model.

6. Discussion and conclusion

The linear regression models are applied for the prediction of FDCs in ungauged basins due to the simplicity in application and requirement of less data when compared to the non-linear regression models. This allows us to obtain a more robust model even if the final performance may be less optimal. Moreover, it is important to recall that the regressions have been applied to the distance matrices, thus requiring the application of the Mantel test to test the significance of each regressor. The current implementation of the Mantel test works in a linear context, while applicability to non-linear regression is yet to be investigated; therefore the method is applied to the linear models only. The simplicity of the proposed procedure makes it a valuable tool for the assessment of FDCs in an ungauged basin.

The procedure is applied to 124 basins in Northern Italy. The basins used in the analysis present different hydrological behaviors and cover a wide range of descriptors (area, elevation, etc.). The methodology is largely centered around the execution of dissimilarity between FDCs in raw (non-logged) space. The dissimilarity matrix is thus generally less sensitive to low flows. The choice of raw versus transformed space depends on the scope of the work (Ganora et al. 2009). In the present work, Equation (1) to compute dissimilarities is applied on non-logged space because the aim is to provide a full representation of the FDC, in order to evaluate

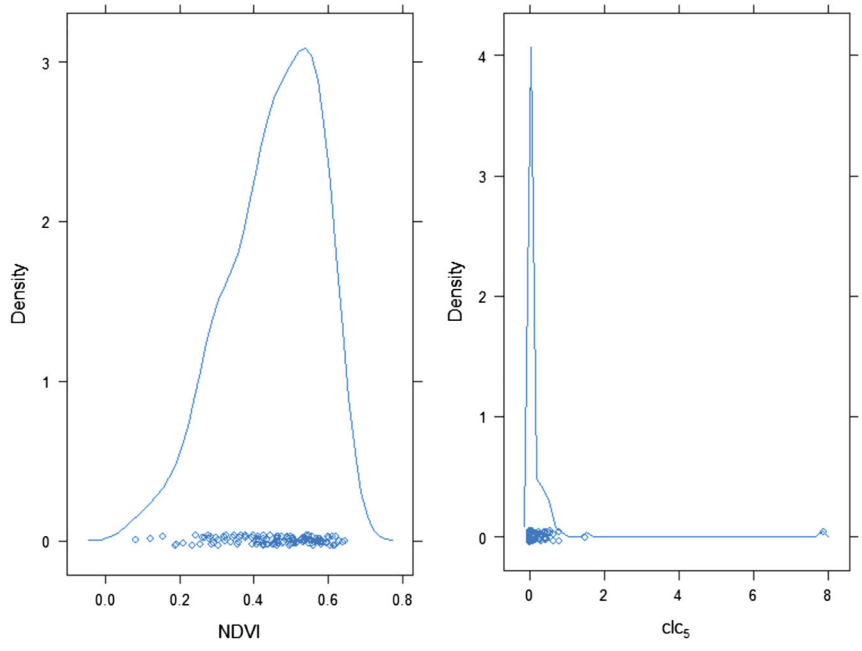

Figure 10. Density plots of some selected descriptors (NDVI and clc_5).

Table 5. Comparison of magnitudes of different errors (ζ) with corresponding standard deviations (δ).

Model	$\zeta_{RMSE}(\delta)$	$\zeta_{NSE}(\delta)$	$\zeta_{D_a}(\delta)$
New method	0.271(0.264)	0.863(0.210)	53.721(37.698)
Geographical method	0.310(0.266)	0.783(0.470)	62.735(38.750)
Parametric method	0.535(0.240)	0.327(0.240)	83.465(35.490)
Single model method	0.301(0.268)	0.798(0.482)	60.112(38.953)

overall volumes rather than just the low-flow tail of the curve.

The discharge distance matrix is linked to the basin descriptors' distance matrices through regression; later the selected descriptors space (obtained from the best regression model) is divided into different clusters and the best models are found for each cluster.

The method also proposed a comparative method called MSP, which provides an opportunity to improve the estimates for RLBs. For an RLB, it is allowed to search the operational model among a set of models with similar global performances but based on the different sets of descriptors with respect to the original model. Alternative models are swapped, searching for the better statistical estimates which can be obtained when the new model better represents the 'location' of the basin in the descriptor space, i.e. the basin is no longer remotely located (or its 'remoteness' is reduced) and the estimation of the design variable is based on a more representative set of donors. This two-step approach allows the global model, which provides the best predictions 'in average', to be more flexible by providing alternative predictions only where the global estimates are less reliable.

The distance-based model proposed here is able to reproduce the unknown FDCs in an efficient way as compared to the geographical distance method and the parametric model. The mean statistical error (in terms of RMSE, NSE and D_a) generated by proposed model is less than the alternative methods (see Table 5). Moreover, the set hypotheses about the ability of proposed procedure (including MSP) to better predict FDCs are further statistically tested. A non-parametric Mann–Whitney U-test is used to evaluate the null hypothesis that the difference in the median of error generated by proposed model and alternative methods is not significant. A probability level of $< .05$ is used to test the null hypothesis, which concluded that the proposed model can reproduce the FDCs by generating lesser magnitude of error.

The present work also covered some of the short falls in previous work done by Ganora et al. (2009): (1) a fixed number of neighbors is not necessarily the best approach in case of RLBs, the model is changed to eliminate remoteness, (2) since the basins are scattered over a wide range of descriptors values, therefore using only a single model for the whole study area is oversimplification. The issue is addressed by dividing the whole study area into smaller clusters and a separate model for each cluster is found, (3) reallocation procedure might be complicated in the case of many clusters whereas no reallocation procedure is required in the proposed methodology. On the contrary, cluster analysis is only done on the descriptor space in the present procedure.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Higher Education Commission, Pakistan [grant number PD (HRDI-UESTPs)/HEC/2012/34].

Notes on contributors

Muhammad Uzair Qamar is an Assistant Professor of Irrigation & Drainage at the University of Agriculture Faisalabad (UAF), Pakistan. His research interests are within the field of hydrology and water resources including hydrological modeling and prediction in ungauged basins (PUB). He has published several peer-reviewed international journal articles in the field of hydrological modeling and PUB.

Daniele Ganora is an Assistant Professor in the Department of Environment, Land and Infrastructure Engineering at Politecnico di Torino, Italy. His research interests are regional models for high and low flows estimation; evaluation of uncertainty; dissimilarity-based method in hydrology; open source software for technological and information transfer. He has published several peer-reviewed international journal articles in the field of hydrological modeling and PUB.

Pierluigi claps is a full professor of Water Engineering since October 2000. He did Ph.D. in Hydrology in 1990. He chairs the Scientific Committee of the HydroAid association and is also a member of the board of the Italian Inter-University Consortium for Hydrology (CINID). He is currently a President of the Gruppo Italiano di Idraulica (Gii). He has authored and co-authored more than 100 papers (25 ISI) in the fields of statistical and watershed hydrology.

Muhammad Azmat is an Assistant Professor of Water Engineering and Management at the National University of Sciences and Technology (NUST), Islamabad, Pakistan. His research interests are within the field of hydrology and water resources including climate change impact assessment and adaptation for the water sector and groundwater assessment and management. Dr Muhammad Azmat has published several peer-reviewed international journal articles in the field of hydrological modeling to climate change adaptation in the water sector.

Dr Muhammad Adnan Shahid has received his doctorate degree in Environmental Engineering from School of Civil and Environmental Engineering, Politecnico di Torino, Italy in 2015. He is currently performing his duties as Assistant Professor (TTS) in Water Management Research Center, University of Agriculture, Faisalabad, Pakistan. His doctorate dissertation mainly focuses on utilization of geoinformatics and hydrologic modeling integrated with open source data for flood management in Pakistan. In addition, he has research expertise in the areas of agricultural water management and irrigation scheduling. Dr Shahid has presented his work in several international conferences and is author of several peer-reviewed publications related to water management, geoinformatics and hydrological modeling.

Dr Rao Arsalan Khushnood received his doctorate degree in the field of Civil (Structural) Engineering from Politecnico di Torino, Italy in 2015. He is a member of Pakistan Engineering Council, Institute of Civil Engineers Uk, American Society of Civil Engineers, and American Concrete Institute. His doctorate research mainly focuses on nano-modificent strategies in high performance self-compacting cementitious systems for the possible improvements in the mechanical behavior, refinement of microstructure and the enhanced shielding effectiveness against electromagnetic interference. He has presented his research work at a significant number of international conferences and authored several peer-reviewed archival journal (ISI

indexed Impact factor) publications. He is currently supervising several MS-Research students in the area of self-healing and nano-modified cementitious matrices.

References

Azmat M, Laio F, Poggi D. 2015. Estimation of water resources availability and mini-hydro productivity in high-altitude scarcely-gauged watershed. Water Resour Manag. 29:5037–5054.

Azmat M, Choi M, Kim T-W, Liaqat UW. 2016. Hydrological modeling to simulate streamflow under changing climate in a scarcely gauged cryosphere catchment. Environ Earth Sci. 75:1–16.

Blöschl G. 2005. On the fundamentals of hydrological sciences. Article 1. In: Anderson MG, Managing Editor. Encyclopedia of Hydrological Sciences. Chichester: J. Wiley & Sons; p. 3–12.

Blöschl G, Sivapalan M, Wagener T, Viglione A, Savenije H, editors. 2013. Runoff prediction in ungauged basins: synthesis across processes, places and scales. Cambridge, UK: Cambridge University Press.

Botter G, Porporato A, Daly E, Rodriguez-Iturbe I, Rinaldo A. 2007. Probabilistic characterization of base flows in river basins: roles of soil, vegetation, and geomorphology. Water Resour Res. 43:W06404. doi:10.1029/2006WR005397

Botter G, Zanardo S, Porporato A, Rodriguez-Iturbe I, Rinaldo A. 2008. Ecohydrological model of flow duration curves and annual minima. Water Resour Res. 44:W08418. doi:10.1029/2008WR006814

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

Castellarin A, Botter G, Hughes DA, Liu S, Ouarda TBMJ, Parajka J, Post D, Sivapalan M, Spence C, Viglione A, Vogel R. 2013. Prediction of flow duration curves in ungauged basins. In: Blöschl G, Sivapalan M, Wagener T, Viglione A, Savenije H, editors. Chp. 7 in runoff prediction in ungauged basins: synthesis across processes, places and scales. Cambridge, UK: Cambridge University Press; p. 135–162.

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

Castellarin A, Galeati G, Brandimarte L, Montanari A, Brath A. 2004a. Regional flow-duration curves: reliability for ungauged basins. Adv Water Resour. 27:953–965. doi:10.1016/j.advwatres.2004.08.005

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

Charrad M, Ghazzali N, Boiteau V, Niknafs A. 2014. Nbclust: an R package for determining the relevant number of clusters in a data set. J Stat Softw. 61(6):1–36. Available from: http://www.jstatsoft.org/v61/i06/

Claps P, Fiorentino M. 1997. Probabilistic flow duration curves for use in environmental planning and management. In: Harmancioglu NB, Alpaslan MN, Ozkul SD, Singh VP, editors. *Integrated approach to environmental data management systems, NATO ASI ser., partnership subser. 2*, vol. 31. Dordrecht: Kluwer; p. 255–266.

Crossland CJ, Kremer HH, Lindeboom HJ, Crossland JIM, Le Tissier MDA. 2005. *Coastal fluxes in the anthropocene: the land–ocean interactions in the coastal zone project of the international geosphere–biosphere programme*. Heidelberg: Springer Verlag.

Farr TG, Rosen PA, Caro E, Crippen R, Duren R, Hensley S, Kobrick M, Paller M, Rodriguez E, Roth L, et al. 2007. The shuttle radar topography mission. Rev Geophys. 45:RG2004. doi:10.1029/2005RG000183

Fennessey NM. 1994. A hydro-climatological model of daily streamflows for the northeast United States [Ph.D. dissertation]. Medford (MA): Tufts University.

Fennessey NM, Vogel RM. 1990. Regional flow-duration curves for ungauged sites in Massachusetts. J Water Resour Plan Manage ASCE. 116:530–549.

Gallo E, Ganora D, Laio F, Masoero A, Claps P. 2013. Atlante dei bacini imbriferi piemontesi (Atlas of river basins in Piemonte) Regione Piemonte.

Ganora D, Claps P, Laio F, Viglione A. 2009. An approach to estimate non-parametric flow duration curves in ungauged basins. Water Resour Res. 45. doi:10.1029/2008WR007472

Ganora D, Gallo E, Laio F, Masoero A, Claps P. 2013. Analisi idrologiche e valutazioni del potenziale idroelettrico dei bacini piemontesi. Progetto RENERFOR Regione Piemonte.

Ganora D, Laio F. 2015. Hydrological applications of the burr distribution: a practical method for the parameter estimation. J Hydrol Eng. doi:10.1061/(ASCE)HE.1943-5584.0001203.

Heerdegen RG, Reich BM. 1974. Unit hydrographs for catchments of different sizes and dissimilar regions. J Hydrol. 22:143–153.

Hosking J, Wallis J. 1997. Regional frequency analysis: an approach based on L-moments. New York: Cambridge University Press.

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

Korn F, Muthukrishnan S. 2000. Influence sets based on reverse nearest neighbor queries. Proceedings of the 2000 ACM SIGMOD international conference on Management of data - SIGMOD '00. doi:10.1145/342009.335415.

Kottegoda NT, Rosso R. 1997. *Statistics, probability, and reliability for civil and environmental engineers*. New York: McGraw-Hill.

Laaha G, Bloschl G. 2006. A comparison of low flow regionalisation methods: catchment grouping. J Hydrol. 323(1–4): 193–214.

Laio F, Ganora D, Claps P, Galeati G. 2011. Spatially smooth regional estimation of the flood frequency curve (with uncertainty). J Hydrol. 408:67–77. doi:10.1016/j.jhydrol.2011.07.022

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

Livneh B, Lettenmaier DP. 2013. Regional parameter estimation for the unified land model. Water Resour Res. 49:100–114. doi:10.1029/2012WR012220

Michele DC, Rosso R. 2002. A multi-level approach to flood frequency regionalization. Hydrol Earth Syst Sci. 6(2): 185–194.

Mimikou M, Gordios J. 1989. Predicting the mean annual flood and flood quantiles for ungauged catchments in Greece. Hydrol Sci J. 34:169–184.

Mohor GS, Rodriguez DA, Tomasella J, Júnior JLS. 2015. Exploratory analyses for the assessment of climate change impacts on the energy production in an Amazon run-of-river hydropower plant. J Hydrol: Reg Stud. 4:41–59.

Montgomery DC, Peck EA, Vining GG. 2001. *Introduction to linear regression analysis*. 3rd ed. New York: John Wiley and Sons, Inc. p. 641.

Müller MF, Dralle DN, Thompson SE. 2014. Analytical model for flow duration curves in seasonally dry climates.

Water Resour Res. 50:5510–5531. doi:10.1002/2014WR015301

Muneepeerakul R, Azaele S, Botter G, Rinaldo A, Rodriguez-Iturbe I. 2010. Daily streamflow analysis based on a twoscaled gamma pulse model. Water Resour Res. 46:W11546. doi:10.1029/2010WR009286

Pandey GR, Nguyen VTV. 1999. A comparative study of regression based methods in regional flood frequency analysis. J Hydrol. 225:92–101 [Pilgrim DH. 19].

Pechlivanidis IG, Jackson B, McMillan H, Gupta H. 2014. Robust informational entropy-based descriptors of flow in catchment hydrology. Hydrol Sci J. doi:10.1080/02626667.2014.983516

Pellicciotti F, Buergi C, Immerzeel W, Konz M, Shrestha A. 2012. Challenges and uncertainties in hydrological modelling of remote Hindu Kush–Karakoram–Himalayan (HKH) basins: suggestions for calibration strategies. Mt Res Dev. 32(1):39–50. doi:10.1659/MRD-JOURNAL-D-11-00092.1

Poff NL, Olden JD, Pepin DM, Bledsoe BP. 2006. Placing global stream flow variability in geographic and geomorphic contexts. River Res Appl. 22:149–166. doi:10.1002/rra.902

Post DA, Jakeman AJ. 1996. Relationships between catchment attributes and hydrological response characteristics in small Australian mountain ash catchments. Hydrol Process. 10:877–892.

Post DA, Jakeman AJ. 1999. Predicting the daily streamflow of ungauged catchments in SE Australia by regionalising the parameters of alumped conceptual rainfall-runoff model. Ecol Model. 123:91–104.

Pumo D, Caracciolo D, Viola F, Noto LV. 2016. Climate change effects on the hydrological regime of small non-perennial river basins. Sci Total Environ. 542(Part A):76–92. doi:10.1016/j.scitoten.2015.10.109

Pumo D, Viola F, La Loggia G, Noto LV. 2014. Annual flow duration curves assessment in ephemeral small basins. J Hydrol. 519(2014):258–270. doi:10.1016/j.jhydrol.2014.07.024

Qamar M, Ganora D, Claps P. 2015. Monthly runoff regime regionalization through dissimilarity-based methods. Water Resour Manage. 29(13):4735–4751.

R Development Core Team. 2013. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Razavi T, Coulibaly P. 2013. Streamflow prediction in ungauged basins: review of regionalization methods. J Hydrol Eng. 18(8):958–975.

Reed DW, Faulkner D, Stewart EJ. 1999. The FORGEX method of rainfall growth estimation – II: description. Hydrol Earth Syst Sci. 3:197–203.

Sefton CEM, Howarth SM. 1998. Relationships between dynamic response characteristics and physical descriptors of catchments in England and Wales. J Hydrol. 211:1–16.

Serinaldi F. 2011. Analytical confidence intervals for index flow duration curves. Water Resour Res. 47:W02542. doi:10.1029/2010WR009408

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

Smakhtin VU. 2001. Low flow hydrology: a review. J Hydrol. 240:147–186.

Smakhtin VU, Weragala N. 2005. An assessment of hydrology and environmental flows in the Walawe river basin Sri Lanka. Working Paper 103. Colombo: International Water Management Institute (IWMI).

Stanoi I, Agrawal D, Abbadi AE. 2000. Reverse nearest neighbor queries for dynamic databases. In ACM SIGMOD workshop on research issues in data mining and knowledge discovery. p. 44–53.

Tung YK, Yeh KC, Yang JC. 1997. Regionalization of unit hydrograph parameters. 1. Comparison of regression analysis techniques. Stoch Hydrol Hydraulics. 11:145–171.

Viglione A. 2007a. nsRFA: non-supervised regional frequency analysis, R package version 0.4–5. Vienna: R Foundation for Statistical Computing. Available from: http://www.r-project.org/

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

Vogel RM, Fennessey NM. 1994. Flow duration curves I: a new interpretation and confidence intervals, ASCE. J Water Resour Plan Manage. 120(4):485–504.

Vogel RM, Fennessey NM. 1995. Flow duration curves II: a review of applications in water resources planning. Water Resour Bull. 31(6):1029–1039.

Wagener T, Sivapalan M, Troch P, Woods R. 2007. Catchment classification and hydrologic similarity. Geogr Compass. 1(4):901–931.

Ward J. 1963. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 58:236–244.

Yokoo Y, Sivapalan M. 2011. Towards reconstruction of the flow duration curve: development of a conceptual framework with a physical basis. Hydrol Earth Syst Sci. 15:2805–2819. doi:10.5194/hess-15-2805-2011

Yusuf MM. 2008. Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves. Hydrol Sci J. 53(4):706–724. doi:10.1623/hysj.53.4.706