Water Resour Manage (2015) 29:4735–4751
DOI 10.1007/s11269-015-1087-7

Monthly Runoff Regime Regionalization Through Dissimilarity-Based Methods

Muhammad Uzair Qamar[1] · **Daniele Ganora**[2] · **Pierluigi Claps**[2]

Received: 10 November 2014 / Accepted: 2 August 2015 /
Published online: 13 August 2015

Abstract A number of procedures can be cited in the literature to perform stream flow prediction in an ungauged basin. Stream flow functions as flow duration curve and flood frequency curves can be obtained by hydrological or statistical models. Also flow regime curves are needed for water resources assessment: they are complex (non monotonic) functions and require special care in the parameterization. Here we propose a dissimilarity-based regionalization model to estimate this particular feature of the stream flow process, as the monthly flow regime. The proposed regional statistical frame work is based on the measure of the dissimilarity (sometimes also referred to as distance) between all the possible pairs of flow regimes available in the region. Each regime is considered as a whole hydrological object and the distance between each pair of regime curves is computed through a suitable metric in a non-parametric way. Dissimilarity values then compose a distance matrix which characterizes the variability of the regime shapes in the region of interest. The prediction of regimes in ungauged basins is obtained by creating corresponding distance matrices of basin features taken among geographic, geomorphologic and climatic attributes, usually referred to as descriptors. Suitable basin descriptors are those whose distance matrices are reasonably correlated to the flow regime distance matrix. This choice allows us to use complex descriptors, like the rainfall regime curve. Identification of the suitable descriptors is performed through an unsupervised procedure based on multiple regressions on distance matrices. Once identified the relations, the candidate descriptors of the ungauged basin can be used to select the most similar gauged basins to use as neighbours for estimation of the required runoff regime. The procedure is applied to a set of 118 basins located in northwestern Italy. The performance of the regional estimation is assessed by means of a cross-validation procedure and through comparison with other parametric regional approaches. In most of the cases, the distance-based model produces better estimates of flow regimes than the "standard" procedure, using only few catchment descriptors, with the advantage of demonstrating the role of complex basin features, as for instance the rainfall regime curve.

✉ Muhammad Uzair Qamar
uzair.qamar@bzu.edu.pk

[1] Department of Agriculture Engineering, Bahauddin Zikariya University, Multan, Pakistan

[2] Politecnico Di Torino, Turin, Italy

Keywords Regression · Regionalization · Distance matrix · Dissimilarity · Hydrological model · Parametric method · Geoghraphical method

1 Introduction

The topic of estimation of flow regimes in an ungauged basin has received extensive research efforts over the last two decades (Blöschl et al. 2013). There are many practical purposes for which prediction of flow regimes is important, as environmental flow requirements, hydropower management, dam storage management for flood mitigation and, of course, irrigation management. In particular, if one considers the monthly flow regime as generally defined, i.e., the curve obtained with the 12 average monthly runoff values in a year, this curve has an important role in design and management of irrigation systems, as seasonality of average runoff is an essential requisite for the deficit assessment. The shape and magnitude of flow regime curves depend on hydroclimatic processes and basin characteristics in a complex way (e.g., Bower and Hannah 2002). The authors above noticed that the basins associated with major aquifers within U.K. are characterized by more stable regimes and the variability in regime shape is a function of seasonal variability and amount of precipitation. They further stated that double peaks are commonly observed in basins associated with large aquifers, whereas prevalence of climatological extremes may result in single regime shape dominating across the entire area.

A number of methods can be cited from literature about flow regime estimation at ungauged sites (e.g., Hrachowitz et al. 2013; Parajka et al. 2013b; Shoaib et al. 2013; Kumar et al. 2015). These methods can be theoretically divided according to Parajka et al. (2013a) into: 1) Process-based methods (e.g., Carrillo et al. 2011) and 2) statistical methods (e.g., Gallart et al. 2008; Samaniego et al. 2010; De Girolamo et al. 2011; Renner and Bernhofer 2011; Archfield et al. 2013). The former are fundamentally based on established physical laws which can capture the underlying dynamics of the watershed. However, they are not suitable for the case of ungauged basins, which is the main goal of the present approach, because they generally require the 'local' calibration of the model parameters. Among statistical methods, is also interesting to cite Olden and Poff (2003), who provided a statistical framework, called *index method*, for the characterization of hydrologic regimes by focusing on the inter-relationships among hydrologic indices. In addition, methods based on geostatistics and proximity concepts have been proposed (e.g., Sauquet et al. 2000, 2008).

Classic regionalization approaches work either on each single monthly value or on a smaller set of representative regime parameters (e.g., Krasovskaia et al. 1994). In the former case an individual regional model is to be defined for each month, which produce simple but cumbersome techniques. On the other hand, using few representative parameters gives the advantage of requiring fewer regional models (i.e., one for each parameter) but thecurve fitting procedure can be complicated. A distance-based method, however, overcomes this choice as it requires only one regional model, defined by a suitable dissimilarity measure, and has no curve fitting requirements. This method is non-parameteric (see Ganora et al. 2009) and aims to estimate of the entire curve as a unique variable.

Another relevant application of the dissimilarity framework is reported by Samaniego et al. (2010) which incorporates copulas to find dissimilarity measures on daily streamflow time series by using three (dis)similarity measure.

Ganora et al. (2009) used regression method to predict flow duration curves by linking descriptors data with hydrological data. To our knowledge no such technique has ever been proposed for flow regimes (non monotonic functions) estimation at ungauged basins, and the magnitude and timing of occurrence of flow regime peaks has never been discussed explicitly earlier.

The dissimilarity-based (or distance-based) method proposed here introduces a new hydrologic metric to consider the dissimilarity between the features of two regime curves. The application of the dissimilarity measure to all the possible combinations of basins, ultimately generates a distance matrix. This distance matrix can then be related to analogous distance matrices computed between other basin characteristics for all basin pairs, with the final aim of using neighbor basins in the space of characteristics to predict the regime curve at an ungauged catchment. This procedure is delineated in the following Sections 1, 2 and 3.

2 Dissimilarity Between Regimes

The dissimilarity-based method we propose starts from the comparison of the flow regimes of a pair of stations. For any two flow regimes belonging to the two gauged basins S and R, constituted by 12 elements each, $\{q_{1,S}, q_{2,S}, \ldots q_{12,S}\}$ and $\{q_{1,R}, q_{2,R}, \ldots q_{12,R}\}$ (i.e., the mean flows of each month), a dissimilarity measure can be defined in different ways. For instance, a function of point to point (magnitudinal) distance between monthly values can be used.

The magnitudinal dissimilarity used by Ganora et al. (2009) reads

$$D_{PtP} = \sum_{i=1}^{12} \left| q_{i,S} - q_{i,R} \right|, \tag{1}$$

where is the monthly mean of the aforementioned stations S or R, D_{PtP} is the point to point difference and i is the index related to the monthly value.

Although Eq. (1) can be applied to flow regimes, it does not account for the possible shifting of peak positioning which is an important feature of flow regimes (see Fig. 1). A more complex definition of distance, accounting for the number and position of local maxima (peaks) and their position can be considered. We thus propose to add to the point-to-point difference D_{PtP} a "lateral distancemeasure" (L_{sp}), which considers the time difference between the occurence of peaks in the two regimes and a"vertical distancemeasure" (V_{sp}), which is the quantitative difference between these peaks. The two measures are then combined in a unique metric to account for all the main features of the regime, i.e., the total distance between two curves is the combination of these three modules (D_{PtP}, L_{sp}, V_{sp}):

$$D_T = D_{PtP} + L_{sp} + V_{sp}. \tag{2}$$

The point to point difference, lateral and vertical separations are sketched in Fig. 1

The vertical distance is assumed as

$$V_{sp} = \left| q_{max,S} - q_{max,R} \right|, \tag{3}$$

where q_{max} is the magnitude of the highest peak discharge at stations S or R.

For estimating the lateral separation, we first need to define the number of peaks in flow regimes. As a starting rule, all the values greater or equal to $0.80 \cdot q_{\max}$ are considered to be

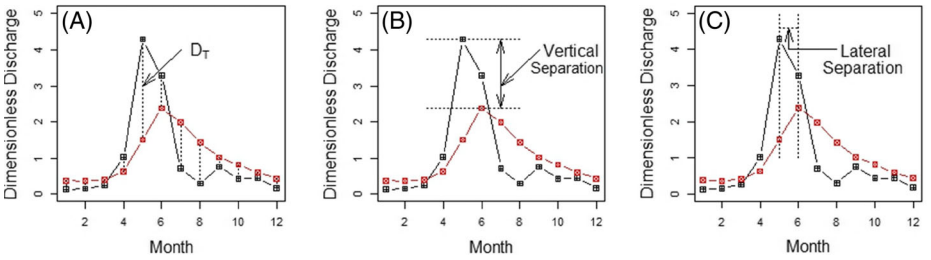

Fig. 1 Distance between flow regimes in the month of May **a** point-to-point distance, **b** vertical seperation of peaks and **c** lateral seperation of peaks

peaks and the lateral separation is computed according to a circular procedure, i.e., the two regimes are shifted towards each other towards the shortest possible span. The regime shifting stops when the moving peak overshadows the peak of the reference station. After a shift the two regimes are referred to be in shifted state, called σ, as opposed to the initial configuration of regimes, called μ. The lateral separation then reads:

$$L_{sp} = \sum\nolimits_i \left| D_{PtP,\mu} - D_{PtP,\sigma_i} \right|, \tag{4}$$

with i the index of the shifted stated.

As an example, L_{sp} is computed by considering station # 9 (see Section 2 for reference) having a peak discharge (S_0) in April and station # 79 having a maximum monthly runoff (R_0) in July, as depicted in Fig. 2. The actual state (μ) of flow regimes at these respective stations refers to the two solid lines. By definition, any of the defined peak (S_0 or R_0) is to be moved towards the other, along the shortest path. Therefore, the movement through these months is going to be backward (July→June→May→April). The process of moving peaks towards each other stops once they are exactly underneath (S_0 or R_3) (see Fig. 2). To obtain the measure of L_{sp} we then need to sum three terms, each one representing an absolute difference as in (4). The first term is the absolute value of the difference between two magnitudinal dissimilarities; $D_{PtP,\mu}$ and $\mu_1 = D_{PtP,\sigma 1}$, the former related to the initial configuration of both curves, and the

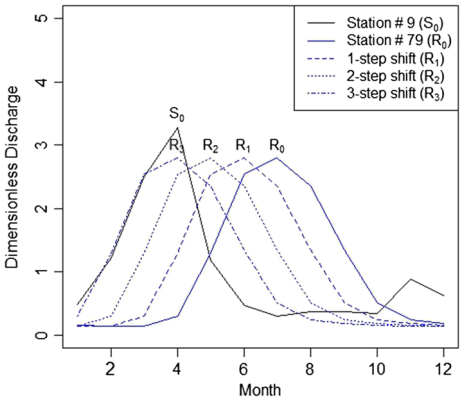

Fig. 2 Compared peaks in actual state (S_0,R_0) and moving R_0 as R_1, R_2,R_3 towards S_0

latter related to the 1-span shift of one of the two regimes. Consequently, the second and the third terms are respectively equal to $D_{PtP,\mu}-D_{PtP,\sigma_2}$ and $D_{PtP,\mu}-D_{PtP,\sigma 3}$.

To understand the difference between using the simple D_{PtP} distance and the comprehensive distance D_T of Eq. (2), a comparison is drawn between two definitions of distances in Fig. 3 based on a set of 118 stations records used in our work (see Section 2 for reference). Their quantitative comparison is done in Table 1,where regimes from three typologies (A, B and C) are put in evidence.

The selection of regime types is operated in three different areas and their differences can be appreciated Fig. 4.

Whereas the points on the bisector line in Fig. 3 are representative of peaks of compared stations that are occuring in the same month (hence no lateral and vertical separation were to be considered). Let us compare a set of randomly selected regimes in blocks A, B and C of the Fig. 3, to understand the difference between D_T and D_{PtP} (or *PtP*). In Fig. 4, the regimes have been actually drawn to further eleborate the difference. The trend (occurence of flow magnitude w.r.t time) of the regimes alongside the time of occurence of peaks is taken into account. In Fig. 4, the regimes in block B are similar to those in A; the reason being small time-scale difference between the occurence of peaks and almost similar trends of regimes being compared in both blocks. By the definition of dissimilarity, the distance of both these blocks should somehow be similar. On the contrary, D_{PtP} distance changes dramatically from A to B but D_T remains consistent. A more simpler case is described in block C, where besides being more similar in C(i) than in C(ii), D_{PtP} counts larger difference between regimes in former and less in latter case. Whereas, D_T reproduces seemingly more meaninful translation of the results as shown in Table 1.

3 Study Area

A dataset of time series from 118 stations in Northwestern Italy was considered for the application. Records have a length varying from a minimum of 5 to a maximum of 52 years, with a mean value of 12 years; the runoff data was extracted from the publications of the former Italian Hydrographic Service, extended with more recent measurements provided by the Regional Environmental Agency (ARPA) of the Piemonte Region. Original data are at the

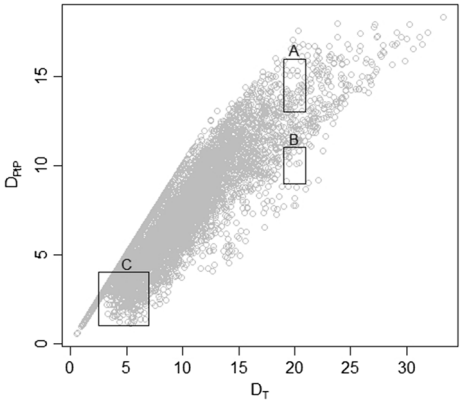

Fig. 3 Comparison between Magnitudinal distance method and the newly developed distance. The points on the bisector are representative of cases in which the peaks of the station pairs are occuring in the same month (hence no lateral and vertical separation were to be considered)

Table 1 Qaulitative comparison of absolute distance method and new method

Stations	Area (km^2)	Mean elevation (m.a.s.l)	Region in fig (3)	D_T	D_{PtP}
84, 89	250, 949	2082, 520	A	20.37	14.23
15, 27	2594, 581	440, 1739	B	20.20	10.09
71, 43	7685, 43	1571, 1655	C	3.01	3.01
71, 41	7685, 1322	1571, 1899	D	6.39	2.80

daily scale and have been aggregated at the monthly scale for the purpose of this study. The data was made dimensionless by normalizing with the global average monthly runoff for each station.

For each of the considered basins a number of climatic, geographic and morphological variables, referred to as descriptors, were taken from the database developed by Ganora et al. (2013) for the region of interest. The catchment area of the considered basins ranges between 22 and 7983 km^2, and their average elevation ranges from 494 to 2694 m.a.s.l. Morphometric characteristics of the basin are based on the NASA Shuttle Radar Topography Mission (SRTM) (Farr et al. 2007) digital terrain model (pre-processed to a 100 m cell grid) using automatic GIS procedures under the Geographic Resources Analysis Support System (GRASS) GIS environment. Climatic, vegetation and land use descriptors were obtained by properly clipping thematic maps available for the area of interest.

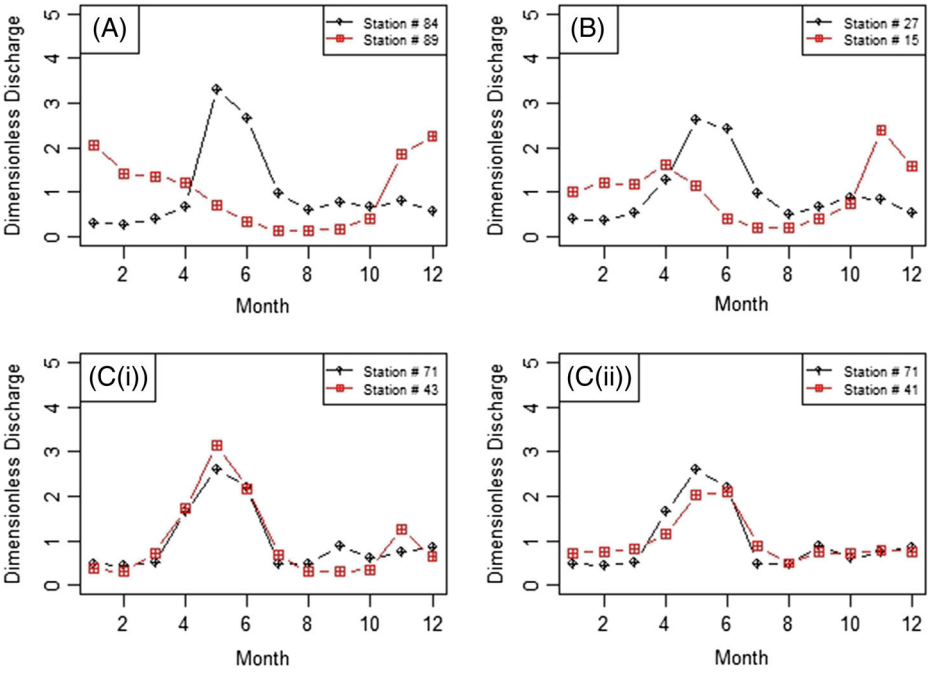

Fig. 4 Examples of regime couples put in comparison so to allow for sensitivity check for dissimilarity methods at various stations

4 Regional Model

The implementation of the regional procedure for regime estimation at ungauged sites is based on the idea that similar hydrological behavior is related to basin similarity in a subset of descriptors. In a Nearest Neighbor approach, similar basins are usually pooled together by proximity in the descriptors space and the average value of hydrological properties, e.g., the flow regime, is taken as valid for the whole group.

Based on the definitions of dissimilarities given in Section 1, one is expected to find low dissimilarity values on descriptors for the basins with "similar" hydrological properties. The simplest descriptors are basin elevation, basin area etc. and the dissimilarity can be computed simply as the absolute difference of the values. When the descriptor is represented by a monotonic function (as the hypsographic curve) the dissimilarity can be computed as the point-to-point distance as in Eq. (1). For more complex descriptors (in this case the rainfall regimes) the D_T dissimilarity is appropriate. Computation of these distances leads to a descriptors dissimilarity matrices.

Only a small subset of descriptors is expected to be representative of the hydrological variability. As there is no prior information about this subset, it is defined through a statistical procedure which looks for the descriptor distance matrix that displays the highest correlation with the distance matrix of the hydrological regime.

The correlation between distance matrices is investigated through the Mantel test (Mantel and Valand 1970). In its simple version, it is used to evaluate the significance of the linear correlation between two distance matrices.

The relation between the discharge distance matrix, defined as M_H, and various combinations of the distance matrices of descriptors (M_D) is in general more interesting than the relationship with one single descriptor. To evaluate this kind of multiple relationship, a linear multiregressive approach has been adopted. We started considering a simple linear model,

$$M_H = \beta_0 + \beta_1 (M_D)_1 + \cdots \beta_n (M_D)_p + \varepsilon, \tag{5}$$

with p as number of descriptors selected among the whole set of available characteristics, β_i as the generic regression coefficient and ε is the residual element of matrices, "unpacked" to vectors as described by Lichstein (2007). In this case, the Mantel test is extended to multiple linear regression models as Eq. (5) as described by Lichstein (2007), who gives details about test implementation. Smouse et al. (1986) also provide useful information for the extension of the simple Mantel test.

Several combinations of models were investigated using linear regression. They were built using all the possible combinations of descriptors distance matrices, based on one to three descriptors at a time. The regressions were first tested for significance with the multiple Mantel test, with a significance level of 0.05. Models passing the Mantel test were then ranked according to the adjusted coefficient of determination defined as (e.g., Kottegoda and Rosso 1997):

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

In the above Eq. (6), p stands for the number of descriptors, n is the total number of basins and R^2 defines the standard coefficient of determination, which alongside

regression coefficients was computed in a standard way, defined by Legendre et al. (1994). As the distances inside a distance matrix are not mutually independent, it is advisable to use all the n values instead of classical $^{n(n-1)}/_{2}$ values. Furthermore, a test against multicollinearity has been performed in order to exclude variables with redundant information in the descriptors.

The R^2_{adj} values observed with distance matrices of regression models are very low (always jaunted between 0.20 and 0.55), although the results are significant, statistically. Which is to say that regressions are only used to select dominating descriptors and not for any direct estimation. This statistic is used to rank the models, but cannot be used to quantify the variance explained by the linear model as in classic regresions due to the mutual correlation of the values in the distance matrices. Besides R^2_{adj} it is of great importance to investigate also the behavior of the residuals along the regression line and its development with time (HP Training Module 2002), which is very difficult to interpret.

To check the quality of model output, we used a cross validation procedure. In this procedure, one station at a time is considered ungauged and its data (hydrological and descriptors) are removed from the database. Afterwards, the models are recalibrated and the 'unknown' flow regime is estimated and compared to the real one. The full scale model validation can be extremely time consuming, depending on the size of dataset and on the complexity of the model, in terms of number of descriptors.

In our work, to reduce the computational burden, the regression models having good R^2 values, filtered through mantel test and VIF test are used to execute the regional regimes and that executed regime is then compared with the empirical regime in D_T space. This means that the error ς, defined as the dissimilarity D_T computed between the empirical and the estimated regime curve is computed for every single station. The model producing least overall error (Δ), between actual and regional regimes was selected, defined as $\Delta = \frac{1}{n}\sum\varsigma$ where n is the number of stations.

The proposed methodology of distance-based measurement was carried out in the R statistical environment (R Development Core Team 2007), desegregated for Mantel test and Multivariate Regression Analysis in nsRFA package (Viglione 2007).

Once the distance-based model is estimated, we find the distance matrices of descriptors in the selected model according to the type of the descriptors (scalar or monotonic). After normalizing them by average distance and then summing them up to find the single representative distance matrix for finding the nearest neighbors (NN) of ungauged basin; considering the minimum value of the distance relative to the stations from the distance matrix of descriptors. The beauty of this technique lies in the ease with which a non monotonic function (complex descriptor) like rainfall was introduced with a scalar descriptor to define an appropriate space for the neighbor selection.

Another important step is to determine the optimum number of neighbors of an ungauged basin. Since too few neighbors resulted in over simplification of the results and in some cases even counter intuitive; whereas, too many neighbors may cause considerable error in the final results. In the present work we used cross-validation procedure to set the number of neighbors and after scrutinizing from 1 to 9, we finally found reasonable results with 5 neighbors.

The best models obtained by one, two and three descriptors are only considered. The best five models constituted by 1, 2 and 3 descriptor(s) against their respective R^2_{adj} and Δ values are ordered according to R^2_{adj} values in Table 2.

Table 2 Models with 1, 2 and 3 descriptors enlisted in the order of R^2_{adj}

Model	Descriptors	Overall error (Δ)	R^2_{adj}
1	Annual NDVI	3.539	0.484
1	Hypsographic Curve	3.862	0.424
1	Mean Basin Elevation	4.067	0.374
1	Max Basin Elevation	3.884	0.216
1	Rainfall Regime	4.044	0.014
2	Fourier Coefficient, Annual NDVI	3.149	0.517
2	Annual NDVI, Rainfall Regime	2.720	0.494
2	Hypsographic Curve, Rainfall Regime	3.018	0.437
2	Mean Basin Elevation, Rainfall Regime	2.940	0.391
2	Land use Index (Non-vegetated area), Rainfall Regime	2.960	0.314
3	Precipitation Intensity Coefficient, Annual NDVI, Rainfall Regime	2.759	0.531
3	Land use Index (Non-vegetated area), Annual NDVI, Rainfall Regime	2.798	0.515
3	Land use Index (Wetlands), Annual NDVI, Rainfall Regime	2.658	0.500
3	Basin Area, Annual NDVI, Rainfall Regime	2.759	0.495
3	Rainfall intensity Duration Curve, Annual NDVI, Rainfall Regime	2.736	0.494

The best results are obtained with three descriptors model (Land use Index (Wetlands), Annual NDVI and Rainfall Regime) due to its higher R^2_{adj} value (0.500) and lower Δ value (2.658) but since the simpler model with two descriptors (Annual NDVI and Rainfall Regime) has comparable R^2_{adj} (0.494) and Δ (2.720) values, therefore considering together three representation of models; the model constituted by Annual NDVI and Rainfall Regime is 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 annual NDVI descriptor is expected to be relevant because of its strong relation to snow accumulation and snowmelt mechanisms. Similarly, the rainfall regime provides a synthetic description of flow pattern. The ranges of some dominating descriptors are enlisted in Table 3.

Table 3 Range of variation of descriptors used by the distance-based model

Descriptors	Maximum	Mean	Minimum
Land use index (Wetlands)	7.890	0.190	0
Rainfall intensity duration curve	37.88	23.40	11.88
Basin area	25640	1330.11	22
Maximum basin elevation	4743	2750	368
Rainfall regime	Regime	Regime	Regime
Mean basin elevation	2682	1323.17	244
Hypsographic curve	Curve	Curve	Curve
Y-coordinate	5129050	4977667	4886350
Land use index (non-vegetated area)	78.68	16.03	0
Annual NDVI	0.644	0.447	0.082
Fourier coefficient	49.563	−8.161	−56.554

The methodology can be summarized in the following steps:

1) Calculate the monthly mean discharge at each station.
2) Identify the variable needed to calculate dissimilarities.
3) Compute dissimilarities between stations by using specified techniques (point-to-point, lateral and vertical).
4) Select best descriptor models by observing least Δ values and Multivariate regression analysis.
5) On the Descriptors space find the NN of missing data station and by using those NN compute the estimation of the regime for that station.

4.1 Alternative Regional Models

4.1.1 Parametric Representation of the Regime

The dissimilarity-based approach was compared with a more traditional regional model based on the parametric representation of the regime curve, which were calibrated on the same set of basins. In contrast to the dissimilarity-based approach which aims at considering the regime as a whole element, here the shape of monthly averaged hydrological regimes is represented by using certain of number of parameters. This parameterization is based on the Fourier harmonic, and its form reads:

$$f(t) = A_0 + A_1 cos\left(\frac{2\pi t}{\tau} + \varphi_1\right) + A_2 cos\left(\frac{4\pi t}{\tau} + \varphi_2\right), \tag{7}$$

where the harmonics represent the 1-year-scale and the 6-months-scale fluctuations of the hydrologic regime. This analytical model to represent the regime has 5 parameters, among which A_0 can be neglected as the mean values is not considered in this work. Phase shifts φ_1 and φ_2 are circular variables so large values may be very close to small values, which on transformation can be sparse apart (e.g., $1^0 * \frac{\pi}{180}$ and $364^0 * \frac{\pi}{180}$). Therefore, in order to estimate them with a regional procedure, it is better to resort to a different representation

$$\begin{aligned} f(t) = A_0 &+ A_1 cos\left(\frac{2\pi}{\tau}t\right).cos(\varphi_1) - A_1 sin\left(\frac{2\pi}{\tau}t\right).sin(\varphi_1) \\ &+ A_2 cos\left(\frac{4\pi}{\tau}t\right).cos(\varphi_2) - A_2 sin\left(\frac{4\pi}{\tau}t\right).sin(\varphi_2), \end{aligned} \tag{8}$$

by separating the variables that do not depend on time t

$$\theta_1 = A_1 cos(\varphi_1); \qquad \theta_2 = A_2 cos(\varphi_2);$$

$$\theta_3 = -A_1 sin(\varphi_1); \qquad \theta_4 = A_2 sin(\varphi_2);$$

and those which depend on t

$$X_1(t) = cos\left(\frac{2\pi}{\tau}t\right); \qquad X_2(t) = cos\left(\frac{4\pi}{\tau}t\right);$$

$$Y_1(t) = sin\left(\frac{2\pi}{\tau}t\right); \qquad Y_2(t) = sin\left(\frac{4\pi}{\tau}t\right);$$

Eq. (8) now reads (neglecting A_0):

$$f(t) = \theta_1.X_1(t) + \theta_2.Y_1(t) + \theta_3.X_2(t) + \theta_4.Y_2(t), \tag{9}$$

whose parameters can be easily fitted to a real dimensionless regime $f(t)$ made of 12 observations by the least squares method (see Fig. 5), where the vectors X_1, X_2, Y_1 and Y_2 are calculated using t=1,2,3……,12 and τ=12

After the fitting procedure of the θ parameters has been extended to all the 118 observed regimes, we proceeded to the regionalization phase. Each parameter θ_j is related to the catchments' descriptors d by a linear model of the form

$$\theta_j = a_0 + a_1.d_1 + a_2.d_2 + \cdots a_n.d_n + \varepsilon, \tag{10}$$

where a_1 are regression coefficients and ε is residual vector. The choice of a suitable regional model is an important step in the estimation of generic parameters at an ungauged basin. Many linear models of the form of Eq. (10) were considered and validated with a Student t test with a significance level of 0.05 followed by a multicollinearity (VIF>5) test and subsequently ordered by their values of R^2_{adj} (e.g., Montgomery et al. 2001).

The leave-one-out validation scheme was used for evaluating the amplitudes and phases of the harmonics and reconstructing the regime. The predicted regime in an ungauged basin is evaluated by combining the basis (X_1,X_2,Y_1 and Y_2) to the estimated θ_j obtained by using the related descriptors. The best models for each θ are;

$$\theta_1 = 4.069*10^{-1} - 6.961*10^{-5}(HypsographicCurve) + 8.795*10H - 4(AverageAspect), \tag{11}$$

$$\theta_2 = 1.298*10^{1} - 1.073*10^{-2}(clc_4) + 2.528*10^{-6}(Basin Latitude), \tag{12}$$

$$\theta_3 = -1.025 + 2.779*10^{-5}(HypsographicCurve) + 1.206*10^{-2}(cn_3), \tag{13}$$

$$\theta_4 = 3.5917 + 0.1473(cn_2) - 0.1684(cn_3), \tag{14}$$

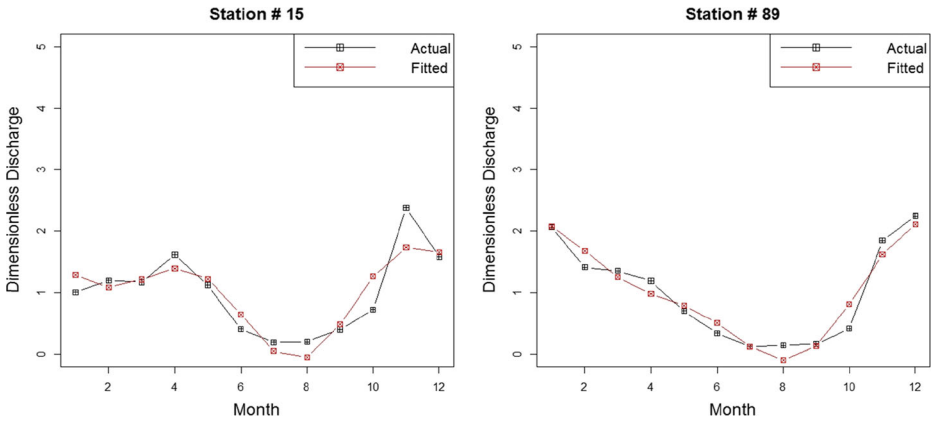

Fig. 5 Fitted regimes over original regimes with parametric models

Table 4 RMSE, NSE and D_T obtained in the various basins using three types of methods

Basin	Area (km^2)	New method			Euclidean			Parametric model		
Stations codes		RMSE	NSE	D_T	RMSE	NSE	D_T	RMSE	NSE	D_T
3	41	0.265	0.918	2.752	0.528	0.675	5.665	0.439	0.775	4.503
6	262	0.382	0.713	4.408	0.483	0.542	8.559	0.462	0.582	5.254
14	127	0.238	0.861	2.301	0.389	0.630	4.144	0.394	0.619	4.692
21	75	0.353	0.820	4.087	0.461	0.692	5.501	0.678	0.334	8.433
28	152	0.145	0.912	1.642	0.388	0.369	7.017	0.373	0.417	6.945
37	106	0.317	0.719	3.246	0.483	0.350	6.323	0.429	0.485	6.864
45	212	0.216	0.900	2.590	0.243	0.873	5.613	0.778	−0.305	13.216
47	102	0.081	0.991	0.814	0.497	0.660	6.185	1.207	−1.008	17.114
48	160	0.251	0.854	4.430	0.445	0.541	5.620	0.770	−0.376	10.089
54	333	0.205	0.605	2.392	0.412	−0.595	4.370	0.448	−0.885	5.409
55	131	0.210	0.729	2.119	0.289	0.486	3.079	0.718	−2.182	9.254
63	838	0.211	0.925	6.855	0.329	0.818	7.378	0.854	−0.228	12.103
70	38	0.157	0.932	1.537	0.390	0.581	4.807	0.532	0.221	7.532
72	25640	0.223	0.664	2.353	0.349	0.173	4.665	0.346	0.186	3.637
82	82	0.198	0.965	2.298	0.525	0.753	6.019	1.387	−0.724	20.012
99	44	0.171	0.750	1.845	0.241	0.501	2.845	0.455	−0.770	4.729
104	249	0.144	0.936	1.302	0.271	0.773	2.904	0.252	0.804	2.913
116	57	0.216	0.899	1.669	0.299	0.807	2.286	0.315	0.786	3.585

Where *clc* and *cn* are corine land cover and soil curve number respectively (for details see Ganora et al. 2013). The error measurement between predicted and actual regimes was obtained by comparing RMSE and NSE values.

4.1.2 Regionalization by Geographical Proximity

The dissimilarity-based approach was also tested against the geoghraphical distance norm which is used to measure the closeness (or dissimilarity) of basins in geoghraphical space. For the sake of simplicity, Euclidean norm was used to find the NN of an ungauged basin. The efficiency of output was tested within a leave-one-out cross-validation scheme.

5 Results and Comparison

The three regional procedures presented in Section 2 provide three different ways to estimate the dimensionless montly regime at ungauged sites. All the methods have been extensively

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

Model	ζ (RMSE,δ)	ζ (NSE,δ)	ζ (D_T,δ)
New method	0.230 (0.091)	0.812 (0.293)	2.578 (1.309)
Geographical method	0.280 (0.11)	0.735 (0.346)	3.363 (1.682)
Parametric method	0.500 (0.221)	0.273 (0.595)	6.370 (3.860)

applied to the 118 basin dataset of Italian catchments described above and are compared in the present section.

Among all the possible models ranked by the distance-based approach, the model containing two descriptors, namely annual NDVI and rainfall regime, was selected for its good global performance in cross validation. More descriptors can be used as well to obtain an enhanced estimator, however increasing the number of descriptor might make the model less robust. For the purposes of this work, the use of only two descriptors is shown to be effective, with performances overtaking those of other regional approaches based on two or more descriptors (Table 4).

A proper metric to quantify the quality of fitting is not trivial to find, for the purpose of comparing the different models. Generally, the metrics are used to compare estimated and observed values (single value comparison), whereas, we need to compare a non monotonic function with a special emphasis on the peak discharge position. It's better to use different metrics to see the goodness of fit of each model by observing the fitting quality of models at each station and ultimately globally. We decided to use RMSE, which is one of the most commonly used error index statistics, and D_T since we are also interested in determining peak flow position.

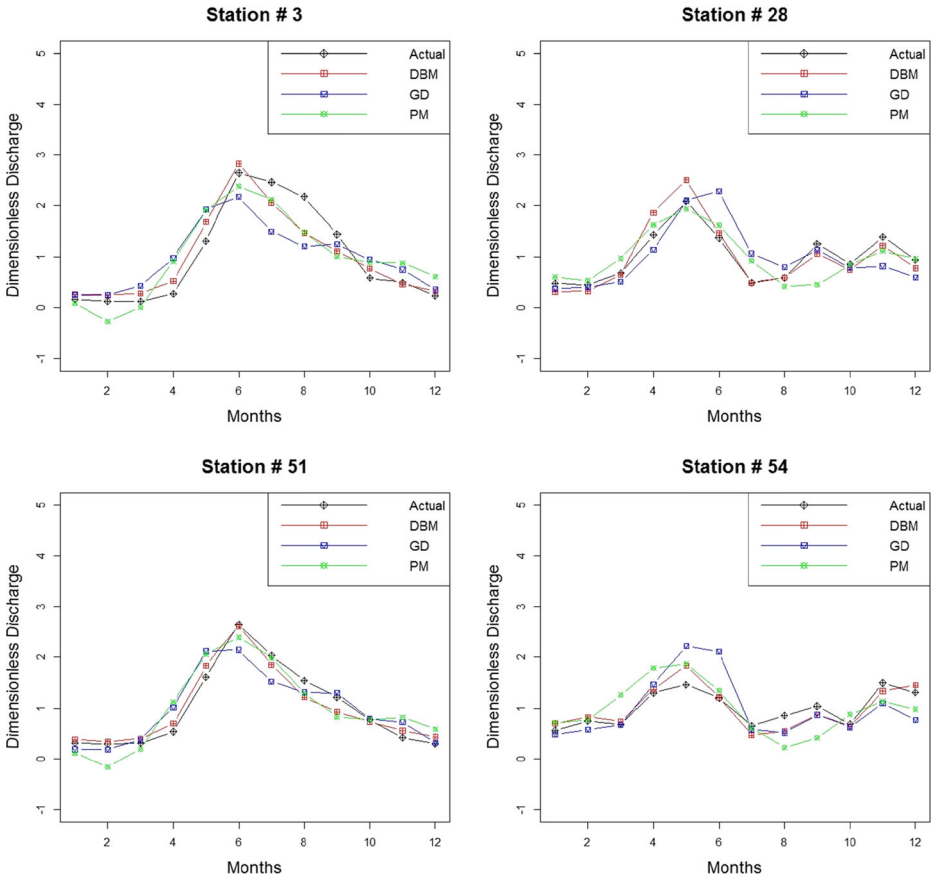

Fig. 6 Comparison between original and simulated regimes at selected stations

On average the distance-based model (*DBM*) has smaller error (ζ) than parametric (*PM*) and geographical proximity (*GM*) as shown in Table 5. Although performances quantified with the D_T metric are expected to favor the distance-based approach, due to peak-shift consideration, the distance-based approach prevails over other models even when RMSE was used for its evaluation.

The newly developed non parametric distance based approach executed, by far, good results compared to those of parametric and geographic proximity models as shown in

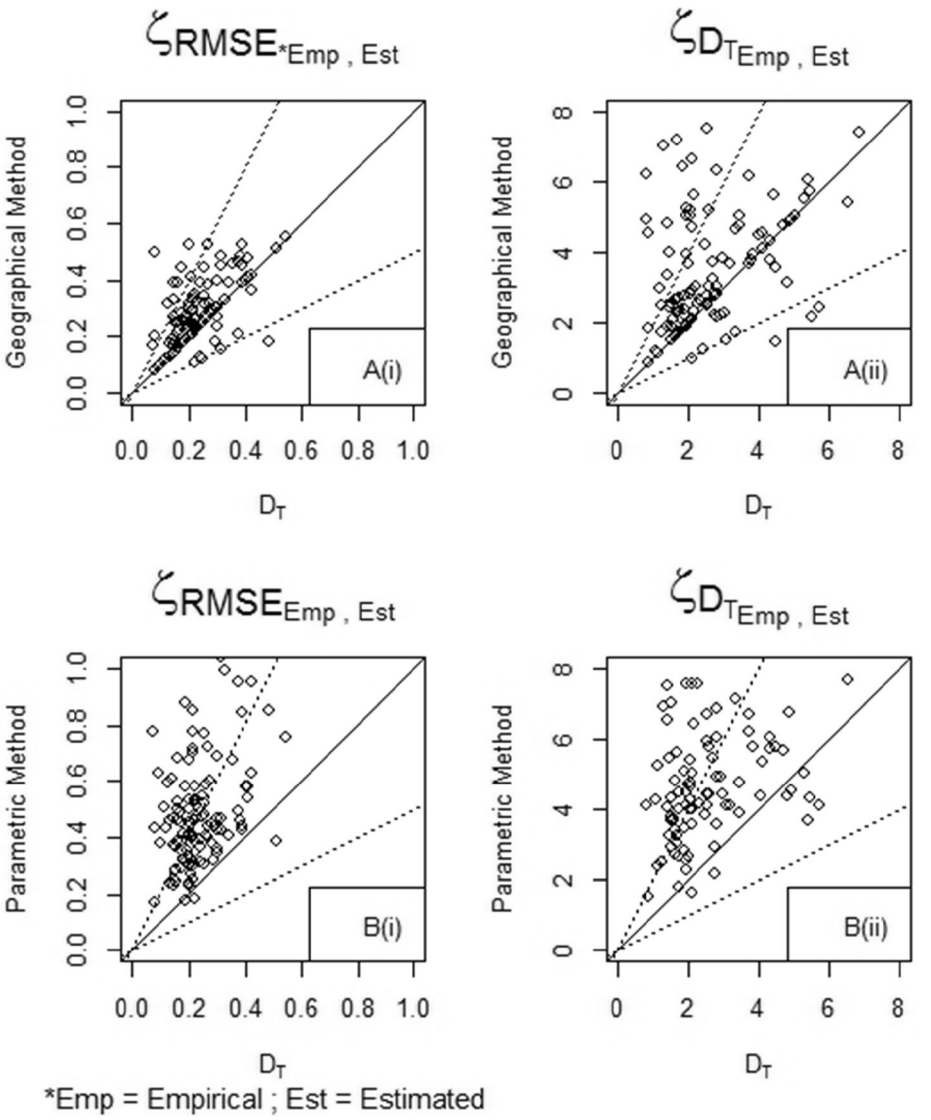

Fig. 7 Error comparison of distance based model in *RMSE* and D_T enviroments compared with **a** the Geograhical method and **b** the Parametric method

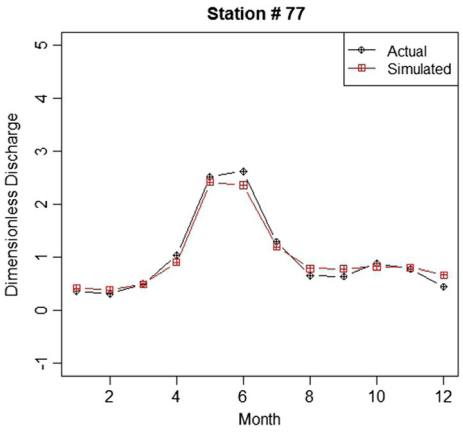

Fig. 8 Estimation of regime in case of flat peak

Fig. 6. The Table 4 illustrates a comparison of D_T, RSME and NSE values among Parametric model, Geographic proximity and Distance-based approach. It was observed that each parametric model was able to execute good results for a certain subset of basins, but not, when tested on the whole dataset. The graphical representation of errors (ζ) obtained in different environments (D_T and RMSE) are shown in Fig. 7, where the distance between the empirical regime and the estimated one, 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-basedmodel are twice the parametric ones and vice versa. Points above the solid line represent regimes better estimated by the distance-based method; points above the top dashed line represent regimes much better estimated by the distance-based method. The total magnitude of error over the entire sample and standard deviation of errors (δ) are enlisted in Table 5.

From these results it can be concluded that the present method led to the most suitable results for flow regimes prediction in most basins with respect to RMSE and D_T. Though the new model performed generally well in all types of catchments, it presented some slight issues of magnitudinal differences between observed and simulated flow regimes for basins with extremely large ($\geq 1000\,km^2$) or small areas ($<100\,km^2$). The model predicted peaks of each regime correctly with slight variation in flat peaks but even in those cases the magnitude of discharge is very close to that of original peak discharge (Fig. 8).

6 Conclusions

The dissimilarity technique between the flow regimes has been revisited in this paper. It has been shown that a good amount of information can be lost by considering, only, magnitude differences (e.g., the monthly-difference of streamflow data) between the flow regimes. While several authors contributed on the identification of the main parameters affecting the shapes of flow regimes, to our knowledge this is the first study which actually tries to integrate all those parameters into a dissimilarity measurement. This measure between regimes is used to account for both the magnitude and the position of the peaks, thus allowing one to quantitatively compare any couple of regimes. This concept is extended to the basin descriptors, so that a dissimilarity index between two sets of basin characteristics can be computed as well.

Information on both flow regime and basin descriptors have been combined to calibrate a regional model: the value of a vegetation index and the average rainfall regime of an ungauged basin are used to identify a set of gauged basins similar to the ungauged one. These are grouped together, and their streamflow records are used to predict the regime at the ungauged site.

The results made available by our distance-based model are comparable and are reasonably better than what we obtained by using other traditional approaches. Moreover, the ability of the modelhere proposed in prediction of complicated annual regimes can be achieved by using only two descriptors.

This approach demonstrates also that is possible to exploit the information of "complex" descriptors, in this case the average rainfall regime, without requiring any kind of parameterization and thus making the prediction procedure easily applicable.

Acknowledgments We would like to thank Dr. Luis Samaniego for his kind help in our work. We also thank two anonymous reviewers for their kind comments on our paper which helped us improving our work. Our work was financially supported by Higher Education Commission of Pakistan under the grant number PD (HRDI-UESTPs)/HEC/2012/34.

References

Archfield SA, Pugliese A, Castellarin A, Skøien JO, Kiang JE (2013) Topological and canonical kriging for design flood prediction in ungauged catchments: an improvement over a traditional regional regression approach? Hydrol Earth Syst Sci 17:1575–1588

Blöschl G et al (eds) (2013) Runoff prediction in ungauged basins. Synthesis across processes, places and scales. Cambridge University Press, Cambridge

Bower D, Hannah DM (2002) Spatial and temporal variability of UK river flow regimes, FRIEND 2002-regional hydrology: bridging the gap between research and practice (Proceedings of Cape Town Conference). IAHS Publ 274:457–464

Carrillo G, Troch PA, Sivapalan M, Wagener T, Harman C, Sawicz K (2011) Catchment classification: hydrological analysis of catchment behavior through process-based modeling along a climate gradient. Hydrol Earth Syst Sci 15:3411–3430

De Girolamo AM, Calabrese A, Lo Porto A, Oueslati O, Pappagallo G, Santese G (2011) Hydrologic regime characterization for a semi-arid watershed, 39–45. In Bodenkultur 62(1–4)

Development Core Team R (2007) R: a language and environment for statistical computing. The R Foundation for Statistical Computing, Vienna

Farr TG et al (2007) The shuttle radar topography mission. Rev Geophys 45:RG2004. doi:10.1029/2005RG000183

Gallart F, Amaxidis Y, Botti P, Cane G, Castillo V, Chapman P, Froebrich J, Garcia-Pintado J, Latron J, Llorens P (2008) Investigating hydrological regimes and processes in a set of catchments with temporary waters in Mediterranean Europe. Hydrol Sci J 53:618–628

Ganora D, Claps P, Laio F, Viglione A (2009) An approach to estimate nonparametric flow duration curves in ungauged basins. Water Resour Res 45:W10418. 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, ISBN:978-88-96046-07-4

HP Training module # SWDP – 37 (2002) How to do hydrological data validation using regression, [Online]. Available:http://www.cwc.gov.in/main/HP/download/37%20How%20to%20do%20hydrological%20data%20validation%20using%20regression.pdf

Hrachowitz M, Savenije HHG, Blöschl G, McDonnell JJ, Sivapalan M, Pomeroy JW, Arheimer B, Blume T, Clark MP, Ehret U, Fenicia F, Freer JE, Gelfan A, Gupta HV, Hughes DA, Hut RW, Montanari A, Pande S, Tetzlaff D, Troch PA, Uhlenbrook S, Wagener T, Winsemius HC, Woods RA, Zehe E, Cudennec C (2013) A decade of predictions in ungauged basins (PUB)—a review. Hydrol Sci J 58(6):1198–1255

Kottegoda NT, Rosso R (1997) Statistics, probability and reliability for civil and environmental engineers, Par 6.2. Mc-Graw-Hill Publishing Company, New York

Krasovskaia I, Arnell NW, Gottschalk L (1994) Flow regimes in northern and western Europe: development and application of procedures for classifying flow regimes. In P Seuna, A Gustard, N. Arnell, W., Cole GA (eds) FRIEND: flow regimes from international experimental and network data (Proc. Braunschweig Conf., October 1993), IAHS Publ. 221, IAHS Press, Wallingford, UK, pp 185–193

Kumar R, Goel NK, Chatterjee C, Nayak PC (2015) Regionalization of watersheds using soft computing techniques. doi: 10.1080/09715010.2009.10514974. http://link.springer.com/article/10.1007/s11269-015-0922-1

Legendre P, Lapointe F, Casgrain P (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: 117–131

Mantel N, Valand RS (1970) A technique of nonparametric multivariate analysis. Biometrics 27:209–220

Montgomery D, Peck E, Vining G (2001) Introduction to linear regression analysis, 3rd edn. John Wiley, New York

Olden JD, Poff NL (2003) Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. River Res Appl 19:101–121. doi:10.1002/rra.700

Parajka J, Andréassian V, Archfield SA, Bárdossy A, Blöschl G, ChiewFHS, DuanQ, GelfanAN, HlavčováK, MerzR, McIntyre N, OudinL, Perrin C, Rogger M, Salinas J, SavenijeH, SkøienJ, Wagener T, ZeheE, Zhang Y (2013a) Predictions of runoff hydrographs in ungauged basins. In Blöschl G, Sivapalan M, Wagener T, Viglione A, Savenije H (eds) Runoff prediction in ungauged basins - synthesis across processes, places and scales, Cambridge University Press (invited), ISBN: 978-1-107-02818-0, 227–360

Parajka J, Viglione A, Rogger M, Salinas JL, Sivapalan M, Blöschl G (2013b) Comparative assessment of predictions in ungauged basins – part 1: runoff-hydrograph studies. Hydrol Earth Syst Sci 17:1783–1795

Renner M, Bernhofer C (2011) Long term variability of the annual hydrological regime and sensitivity to temperature phase shifts in Saxony/Germany. Hydrol Earth Syst Sci 15:1819–1833

Samaniego L, Bardossy A, Kumar R (2010) Streamflow prediction in ungauged catchments using copula-based dissimilarity measures. Water Resour Res 46:W02506. doi:10.1029/2008WR007695

Sauquet E, Gottschalk L, Leblois E (2000) Mapping average annual runoff: a hierarchical approach applying a stochastic interpolation scheme. Hydrol Sci J 45(6):799–815

Sauquet E, Ramos MH, Chapel L, Bernardara P (2008) Streamflow scaling properties: investigating characteristic scales from different statistical approaches. Hydrol Process 22:3462–3475. doi:10.1002/hyp.6952

Shoaib SA, Bárdossy A, Wagener T, Huang Y, Sultana N (2013) A different light in predicting ungauged basins: regionalization approach based on eastern USA catchments. J Civ Eng Archit USA 7(3):364–378, **ISSN1934-7359**

Smouse PE, Long JC, Sokal RR (1986) 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