Hydrology Research

British Hydrological Society

IWA PUBLISHING

Hydrology Research Vol 53 No 8, 1075 doi: 10.2166/nh.2022.022

Regionalization of flow duration curves in Colombia

Carlos Gaviria [a,*] and Fernando Carvajal-Serna [b]

[a] Ingeniero Civil, Magister en Ingeniería de Recursos Hidráulicos, Facultad d eMinas, Universidad Nacional de Colombia sede Medellín, Colombia
[b] Associate Professor, Universidad Nacional de Colombia, Sede Medellín, Facultad de Minas, Departamento de Geociencias y Medio Ambiente, Posgrado en Aprovechamiento de Recursos Hidráulicos, Carrera 80 no 65-223-Medellín-AA 1027-Colombia
*Corresponding author. E-mail: cjgaviriaa@unal.edu.co

CG, 0000-0002-8774-6790; FC-S, 0000-0003-4149-1451

ABSTRACT

It is essential to know the streamflow behavior in hydrological basins for appropriate water resource planning and management. In Colombia, where there is a considerable water resource potential, there is a need to generate hydrological modeling for many ungauged catchments. Thus, this study presents the regionalization of flow duration curves (FDC) in Colombia. Daily flow time series from 655 gauging stations were used to define homogenous hydrological regions, considering geological, topographic, and climatic information. Fifteen hydrological regions were delimited by cluster analysis using the K-means algorithm, all of which exhibited high spatial heterogeneity. Multiple linear regressions were used to estimate characteristic dimensionless flows as a function of each basin's attributes. A set of equations that allow the reconstruction of simulated dimensionless FDC for each cluster was determined, and regression (R^2) values of 0.5–0.9 were obtained. The percentage error of the mean, maximum, and minimum discharge of the simulated FDC compared with observed values were approximately 9, 30, and 50%, respectively.

Key words: flow duration curves hydrological cluster, K-means, lineal multiple regression, scarce information hydrology

HIGHLIGHTS

- Regional equations that allow the estimation of flow regime in Colombia.
- Definition of hydrological homogeneous regions in Colombia.
- Hydrological data and parameter analysis for Colombia.
- More regionalization parameters were included and analyzed.
- Linear multiple regresions were performed with more than two independent variables.

1. INTRODUCTION

Flow estimation in ungauged catchments is an essential task for the study, planning, and management of water resources worldwide and remains a major challenge for the hydrological community (Sivapalan 2003). It is necessary to know the flow regime behavior estimations on catchments or territories lacking information (Mesa Sánchez *et al.* 2003). Flow duration curves (FDC) can be applied to summarize flow regimes (Foster 1933) using the relationship between their magnitude and frequency (Vogel & Fennessey 1994). Many engineering and environmental planning applications of FDC exist (Castellarin *et al.* 2007): for example, the analysis of maximum, average, and minimum flows (Blöschl 2005) and the estimation of environmental flow and water supply at points of interest within the framework of water planning and management (GWP 2008; IDEAM & MinAmbiente 2015). There are different methodologies for the parameterization of the FDC, one of them is based on regression models as in this study (Wagener *et al.* 2013) and others are based on physical models supported by a probability distribution (Doulatyari *et al.* 2015).

Colombia has a considerable water resource potential (IDEAM 2008), which generates the need for knowledge about flow regime for appropriate designs, execution, and operation of projected small and large hydroelectric power plants, intake water or potable water, agriculture, flood control, recreation, etc. (UPME-PPUJ 2015). The main motivations for developing this

work are the lack of information and the need to adequately characterize the flow regime in Colombia for those mentioned above.

The main objective of this work is to present a method that permits the streamflow approximations in ungauged catchments using FDC estimation for different hydrological regions in Colombia (Olden *et al.* 2012). The information aggrupation technique defines these hydrological regions on the basis of the K-means cluster methodology (Hartigan 1975). Multiple linear regression (MLR) analysis is proposed to estimate the FDC on each cluster, interpolating dimensionless flow estimations of different characteristic percentiles (Li *et al.* 2010). Different methodology and approximation of FDC are presented in this study (Vogel & Fennessey 1994; Castellarin 2014).

This research considers more catchment attributes, including geological (Musiake *et al.* 1975), geomorphological (Beven 2012), climatic (Perez *et al.* 2019), topographical (Wood *et al.* 1990), landscape (Winter 2001), and vegetation descriptors (Burt & Swank 1992). Correlations with more than two variables were used to estimate the percentiles of characteristic flow and more criteria for hydrological regions per definition (Mohamoud 2008). Furthermore, a comparison between cluster flow estimation results and traditional subregions aggrupation flow estimation is performed.

This work considers different aspects that improve the FDC estimation in Colombia: The proposed estimation of the FDC takes into account the high spatial and temporal variability in Colombia considering a regionalization strategy for the hydrographic zones using morphometric criteria, hydrological information, and the K-means method.

This work presents 15 homogeneous hydrological regions instead of the 5 traditional regions (Caribbean, Pacific, Andean, Amazon, and Orinoquia) which allows considering areas that have different microclimates inside them and even speculate tele-connections between areas or basins in different parts of the country. Observed FDC were estimated using as much information as possible after depuration of inconsistencies or time series of less than 15 years of length.

Also, FDC were estimated using a piecewise fit strategy to represent, in the best way, the minimum, mean, and maximum flow regime. Another methodological contribution consists of adding more than two variables to the equations built from the MLR method and obtaining equations that relate up to five independent variables. This paper is presented as follows: data and methodology, results, discussion, and conclusions.

2. DATA AND METHODOLOGY

2.1. Data collection

A total of 655 daily flow time series data were collected from measurement stations (IDEAM 2014). The length of data was from 1940 to 2015. In determining the gauging stations, the measurement stations were distributed in 1,141,748 km², corresponding to the surface of the Colombian national territory located in the northern region of South America. Time series corresponding to gauging stations located in river branches with multiple channels, series with less than 15 years of daily records, and those detected as inconsistent were removed from the analysis.

Topographic data were collected from the shuttle radar topographic mission model with a 90 m resolution (NASA & Watkins 2014) for all Colombian territories. Land cover and soil type information were taken from the Instituto Geográfico Agustín Codazziigac (IGAC) through platform *Sistema De Información Geográfica Para La Planeación Y El Ordenamiento Territorial* (IGAC 2014). Northern South American monthly precipitation reanalysis from Hurtado (Hurtado & Mesa 2014) was used to estimate the mean and maximum precipitation on each catchment, whereas Cenicafé equations were used to calculate potential evapotranspiration and mean surface air temperature (Jaramillo 1989; Barco & Cuartas 1998; Chaves & Jaramillo 1998). Figure 1 shows the study zone and gauging stations.

2.2. Hydrological clustering

After drawing the hydrographical basin corresponding to selected gauging stations (Figure 1(b)), each gauging station was assigned the set of attributes and climate to landscape descriptors to perform hydrological clusters using the K-means algorithm (Álvarez *et al.* 2011; Wilks 2011). Table 1 shows these sets. The entire national territory was divided into subbasins or hydrological units to extrapolate and spatialize the cluster results. Sensitivity analysis with the mean Euclidian distance to each cluster centroid was conducted to define the optimum number of basins groups (García *et al.* 2017).

2.3. FDC estimation using a simple linear regression model

From observed FDC built including all observed data available at the calibration stations (period-of-record FDC), the distribution of gauged basins in clusters (Sauquet & Catalogne 2011), climate and landscape attributes, and the values of different

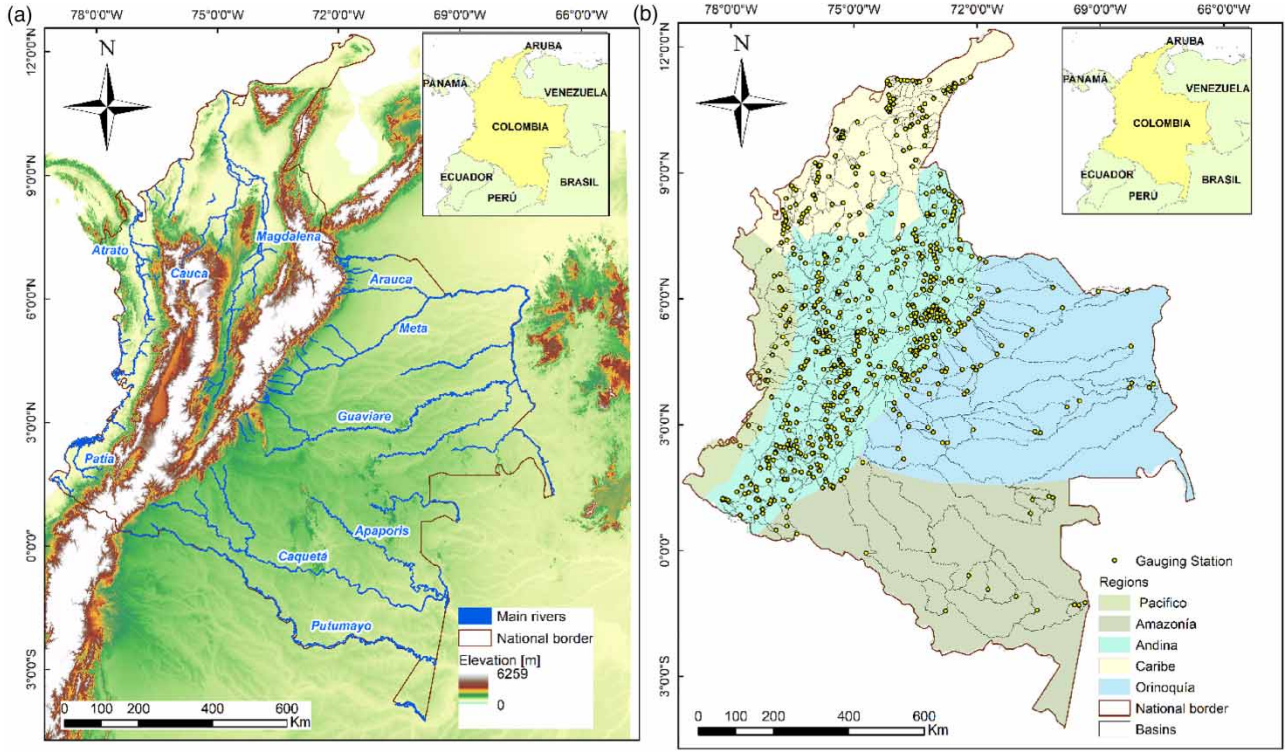

Figure 1 | (a) Study zone. (b) Gauging station locations. Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/nh.2022.022.

characteristic flows were estimated to generate daily, synthetic, and regional duration curves (Razavi & Coulibaly 2013). Characteristic flow percentiles Q_p^* chosen were similar to those of Mohamoud (2008) and Salazar Oliveros (2016): Q_{100}, Q_{90}, Q_{80}, Q_{70}, Q_{60}, Q_{50}, Q_{40}, Q_{35}, Q_{30}, Q_{20}, Q_{10}, Q_5, Q_1, $Q_{0.5}$, and $Q_{0.1}$. Characteristic flows Q_p were normalized using the relationship with the average streamflow $\bar{Q}$, resulting in dimensionless characteristic flows as follows: $Q_p^* = Q_p/\bar{Q}$. Due to the magnitude relationship between the FDC and the drainage area or the average flow, it is necessary to standardize, in this case removing dimension by the relation on the average flow of each series. Regionalization tests were carried out based on a standardization or normalization with respect to the drainage area of each basin; however, the results were not the best.

MLR (Anderson 1958) differs from simple linear regression in analyzing the influence of one but several explanatory variables X on a dependent variable Y (Rojo Abuín 2007). The general form for MLR is given by Equation (1).

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n + \beta \tag{1}$$

where y represents each one of the dimensional flow percentiles Q_p^*, α_i represents the value of each term coefficient, x_i represents each regression attribute, and β is the ordinate axis intersect or independent term.

Here, the MLR was used in its potential form (Equation (2)). Thus, logarithmic transformation was applied to estimate linear regression parameters.

$$\begin{aligned} \log(y) &= \alpha_1 \log(x_1) + \alpha_2 \log(x_2) + \ldots + \alpha_n \log(x_n) + \log(\beta) \\ \log(y) &= \log(\beta\, x_1^{\alpha_1}\, x_2^{\alpha_2} \ldots x_n^{\alpha_n}) \\ 10^{\log(y)} &= 10^{\log}(\beta\, x_1^{\alpha_1}\, x_2^{\alpha_2} \ldots x_n^{\alpha_n}) \\ y &= \beta\, x_1^{\alpha_1}\, x_2^{\alpha_2} \ldots x_n^{\alpha_n} \end{aligned} \tag{2}$$

Table 1 | Set of attributes and landscape to climate descriptors

Variable	Units	Abbreviation
Basin drainage area	km²	DreA
Basin perimeter	km	Perim
Graveluis compactness coefficient	m/m	Comp
Agricultural land percentage	%	%Agr
Forest land percentage	%	%For
Urban land percentage	%	%Urb
Tectonic fault density	(km/km^2)	FDen
Drainage density	(km/km^2)	DDen
Slime percentage	%	%Lo
Sand percentage	%	%San
Clay percentage	%	%Cl
Mean annual potential evapotranspiration	mm/year	PET
Maximum elevation	m	MaxE
Mean elevation	m	Emean
Minimum elevation	m	MinE
Basin unevenness	m	BUne
Mainstream channel length	km	MSCL
Average basin slope	%	Sl
Maximum monthly precipitation	mm/month	Pmax
Mean monthly precipitation	mm/month	Pm
Mean surface temperature	°C	Tm
Hypsometric curve percentile 10	%	H_{10}
Hypsometric curve percentile 25	%	H_{25}
Hypsometric curve percentile 50	%	H_{50}
Hypsometric curve percentile 75	%	H_{75}

Equation (2) represents the potential form of MLR, and it is a multiplicative of n in terms x_i, raised to its respective α_i exponent and a general coefficient β. Contrary to Equation (1), coefficient β is the logarithm of the independent term, whereas exponents α_i are the same, and x_i and y values are the logarithm of the original matrices. The potential equation was adequate to represent natural processes. In this case, discharges result from the interaction of the multiple product variables, such as terrain slope, soil covers, drainage area, and catchment perimeter.

Dimensionless flow estimations were also performed for the five traditionally established subregions in Colombia: Caribe, Magdalena, Orinoquia, Amazonia, and Pacific (Salazar-Holguín, 2013). These regions were delimited in this way because of their similarity in topographical, climatic, and geomorphological aspects on a large scale, as well as because of their geographical position in the national territory. Figure 1(b) shows these regions. Results were contrasted between estimation from clusters and traditionally established subregions, and the hypothesis indicated that estimation applied to clusters should give better approximations to observed flows (Swain & Patra 2017).

Statistical model (*Regress*) was performed to select the combination of variables that presents the highest value of determination coefficient R^2 on each equation (Mohamoud 2008), which establishes how good the estimations of Q_p^* were (regression model output) (Steel & Torrie 1960). Fifteen matrices (one by each cluster) were conformed to dimensions $n \times 25$, where n *represents* the number of selected gauging stations and 25 represents the attributes in Table 1; this matrix group was named X. The dependent variable Y was compounded using 15 matrices (number of clusters) with dimensions $n \times 15$ (number of characteristics percentiles) and dimensionless characteristic streamflow Q_p^*. However, data were not standardized in both cases. The regression analysis was very useful, and this permits to obtain dimensionless FDC for each homogeneous region and its validation exercise, respectively.

From each hydrological cluster and subregion, a flow time series and its respective catchment were randomly selected to validate the estimation results. These stations were excluded from the calibration process. Tables 2 and 3 show the validation station for clusters and subregions, respectively.

3. RESULTS

3.1. Hydrological clustering

The K-means algorithm was run several times with data bank numerically standardized, varying the parameter Number of Cluster. Figure 2 presents the results of Euclidian mean distance variation to the cluster's centroid. Fifteen clusters were defined as an optimum number to work with (García *et al.* 2017).

The 655 hydrological basins were grouped into 15 clusters using the K-means algorithm. Table 4 shows a summary of clustering results.

Ungauged basins were grouped with the same criterion and procedure as gauged ones. Figure 12 shows the result of this grouping.

3.2. FCD estimations

MLR was conducted for FDC estimation with two independent variables on each equation. R^2 average coefficient for each cluster was comparatively qualified by its value as follows: *Poor* if R^2 coefficient is lower than 0.3, *Fair* if it is between 0.3 and

Table 2 | Validation stations for clusters

Cluster	Station code	Region
1	11077020	Caribe
2	23097040	Magdalena
3	35017070	Orinoquia
4	21227010	Magdalena
5	13017010	Caribe
6	35027020	Orinoquía
7	21207960	Magdalena
8	51027020	Pacific
9	42067010	Amazonia
10	15017010	Caribe
11	32077100	Orinoquia
12	23057010	Magdalena
13	21017020	Magdalena
14	35027150	Orinoquia
15	21197030	Magdalena

Table 3 | Validation station for subregions

Region	Station code
Caribe	13047040
Magdalena	21147080
Orinoquía	35087010
Amazonía	44117010
Pacífico	52027030

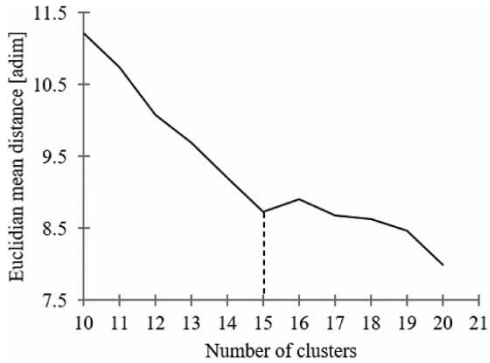

Figure 2 | Sensitivity analysis – definition number of clusters.

Table 4 | Clustering results

Cluster	Number of units	Predominant subregion
1	53	Magdalena
2	21	Magdalena
3	12	Magdalena and Orinoquia
4	82	Magdalena and Caribe
5	38	Caribe and Orinoquia
6	29	Magdalena and Orinoquia
7	16	Magdalena
8	62	Magdalena
9	26	Orinoquia and Amazonia
10	57	Magdalena
11	38	Caribe and Pacific
12	42	Magdalena and Amazonia
13	74	Magdalena
14	65	Magdalena
15	40	Orinoquia
	Total: 655	

0.4, *Good* if it is between 0.4 and 0.6, and *very good* if R^2 value is greater than 0.6. The same results for traditional subregions were estimated.

Generally, higher R^2 values were obtained in cluster FDC estimations compared with those obtained in traditional subregions estimations.

MLR was recalculated to improve flow estimation in different clusters and percentiles in which qualification was fair or poor, adding more independent variables to the equations. Higher R^2 values were observed, with an average increase from 0.45 to 0.54. Table 5 presents a summary of the results.

Figure 3 shows the average R^2 coefficient variation between magnitudes (flow percentiles) for estimation equations. Unlike a low R^2 value in Q_{20} and a decrease toward minimum flows, there is no significant trend in Figure 3. Figure 4 shows the percentage of participation of each attribute on the regression equations. Among influential attributes are the percentage of forest, percentage of agriculture, mainstream channel length, and percentage of urban areas.

A test was performed on apparently similar clusters: the dimensionless flow regime was estimated with the parameters of cluster 15 applied to cluster 14. The results were very abnormal, in contrast to the results obtained with the corresponding

Table 5 | MLR results – two or more variable equations in the cluster

Cluster	Average R^2	Number of variables	Concept
1	0.46	4	Good
2	0.79	2	Very good
3	0.78	2	Very good
4	0.40	5	Good
5	0.60	2	Good
6	0.53	5	Good
7	0.71	2	Very good
8	0.48	4	Good
9	0.54	3	Good
10	0.49	4	Good
11	0.71	2	Very good
12	0.43	4	Good
13	0.35	5	Fair
14	0.36	5	Fair
15	0.49	5	Good

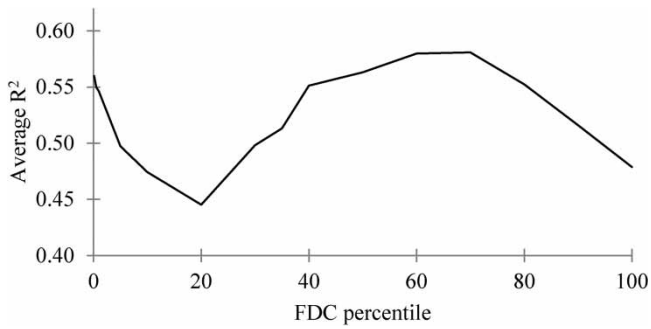

Figure 3 | R^2 variation between percentiles.

parameters. Figures 5–7 show three graphic examples of FDC estimation and interpolation for clusters (3 and 9) and traditional subregions (Magdalena region). Conversely, Figures 8–10 show their respective dispersions with $y = x$ function.

Table 6 shows the linear correlation coefficient (R) and covariance (R^2) between observed and estimated streamflow, and $y = x$ function for clusters, whereas Table 7 shows the results for traditional subregions.

Table 8 shows the mean relative percentage error between observed and estimated flow for cluster grouping estimations, whereas Table 9 shows those of traditional subregions estimations. Figure 11 shows the comparison between validation percentage errors for both regionalization cases. It can be observed that estimations from regions got better average relative errors than cluster grouping estimations.

Dimensionless streamflow estimation equations are shown in the Supplementary Material (Gaviria Arbeláez 2019). Although the equations follow Equation (2) structure, it is highlighted that the combination of attributes shown in each percentile regression has a higher R^2 value. Figure 12 shows the map of the regions with the main results of FDC estimations (validation set), in which the horizontal and vertical axes represent the percentage of exceedance and dimensionless flow, respectively.

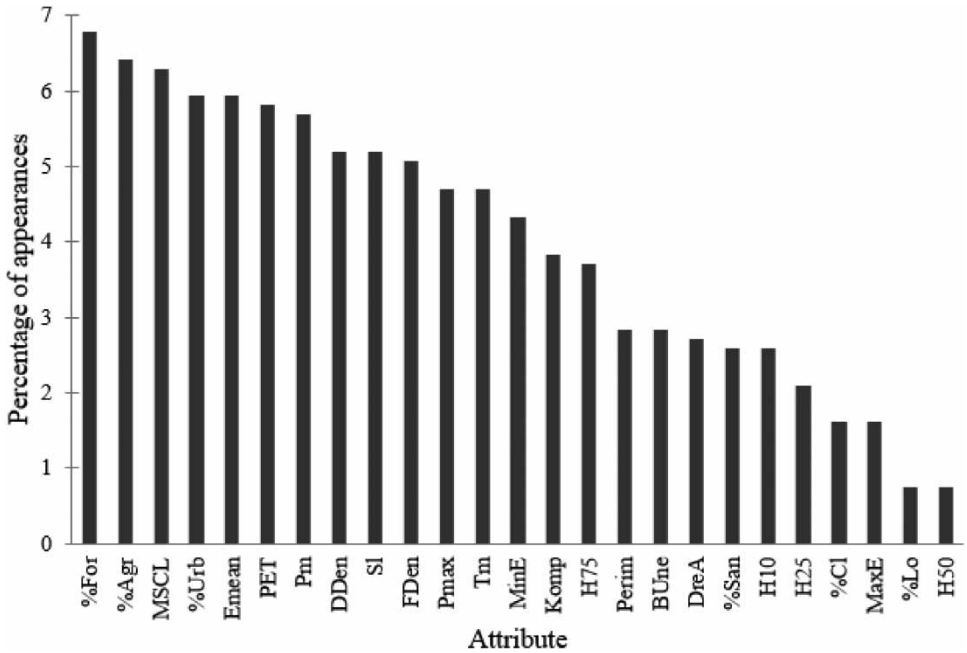

Figure 4 | Percentage of appearances of each attribute.

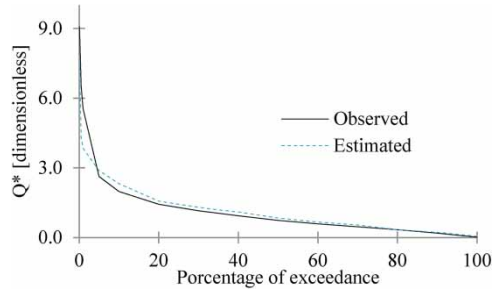

Figure 5 | Example – observed and estimated FDC – Cluster 3 – Orotoy river – drainage area: 167 km^2.

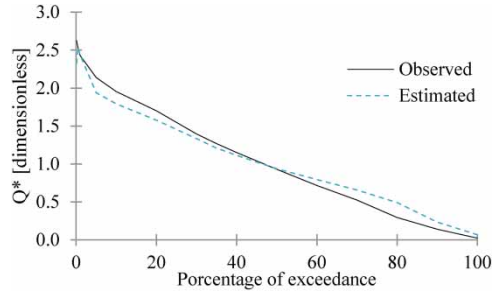

Figure 6 | Example – observed and estimated FDC – Cluster 9 – Vaupés river – drainage area: 17,070 km^2.

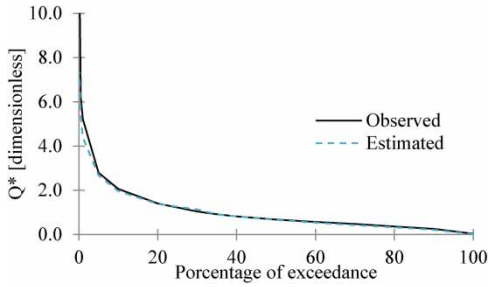

Figure 7 | Example – observed and estimated FDC – Magdalena region – Cabrera river – drainage area: 1,185 km^2.

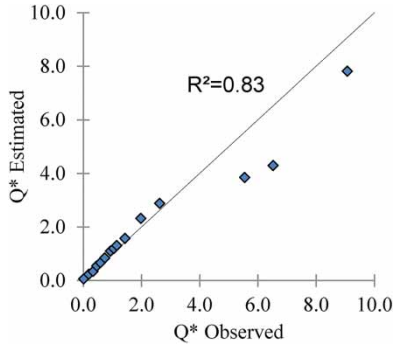

Figure 8 | Dimensionless estimated flow vs. observed dimensionless flow – Cluster 3 – 35017070.

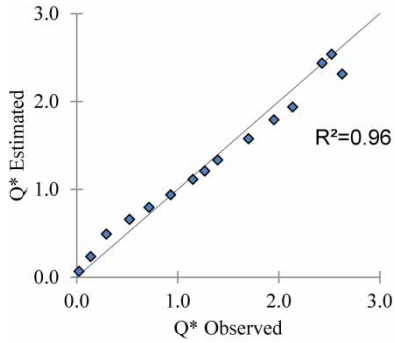

Figure 9 | Dimensionless estimated flow vs. observed dimensionless flow – Cluster 9 – 42067010.

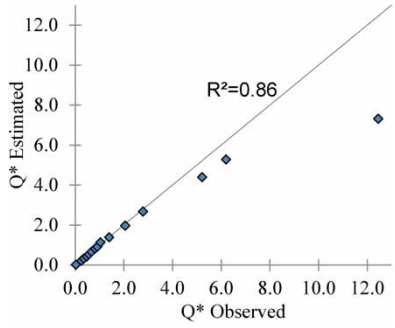

Figure 10 | Dimensionless estimated flow vs. observed dimensionless flow – Magdalena region – 21147080.

Table 6 | R and R^2 values for observed and estimated streamflow vs. $y = x$ on clusters

Cluster	Station	$R-y=x$	$R^2-y=x$
1	11077020	0.86	0.74
2	23097040	0.98	0.97
3	35017070	0.91	0.83
4	21227010	0.79	0.62
5	13017010	0.93	0.86
6	35027020	0.93	0.86
7	21207960	0.80	0.63
8	51027020	0.96	0.92
9	42067010	0.98	0.96
10	15017010	0.82	0.67
11	32077100	0.95	0.90
12	23057010	0.96	0.91
13	21017020	0.91	0.82
14	35027150	0.91	0.82
15	21197030	0.83	0.70

Table 7 | R and R^2 values for observed and estimated streamflow vs. $y = x$ on traditional subregions

Region	Station	$R-y=x$	$R^2-y=x$
Caribe	13047040	0.99	0.98
Magdalena	21147080	0.86	0.75
Orinoquia	35087010	0.94	0.89
Amazonia	44117010	0.94	0.88
Pacific	52027030	0.96	0.92

4. DISCUSSION

High spatial heterogeneity and discontinuities were seen in the cluster regions map (Figure 12). This can be explained by the complexity and orography of Colombian territory, high spatial variability of precipitation, and heterogeneity of land cover and soil type.

The main result is an equation that estimates the dimensionless flow of each characteristic percentile and cluster. Generally, higher correlations were found in cluster regression than in regional equations. Nonetheless, locating a study catchment in a traditional subregion would be simpler than locating it in a regionalization cluster.

The mean percentage error in model validations gave an average of approximately 27%, corresponding to 50, 9, and 26% of minimum, mean, and maximum flow estimations. However, traditional subregions validations generally gave a higher performance, as shown in Figure 11. Most of the streamflow magnitudes gave lower percentage error in regions estimations than in clusters. This lower percentage error was different than expected because the cluster regression equation gave generally higher R^2 values. Also, the regions were validated with basins of different sizes to show that the results fit for a wide range of areas

Estimation behavior varies spatially between clusters. For example, better results were found in clusters 2, 3, 7, and 11, which gave considerably higher R^2 values with respect to the other groups. By contrast, clusters 13 and 14 gave lower R^2 values. This behavior was not due to the number of calibration points or how homogeneous each group was. Nevertheless, even when R^2 values are not so high, it is possible to have satisfactory approximations when the estimated and observed flows are compared. It also depends on the study catchment, which is the case validations of clusters 8, 9, 10, and 11.

Table 8 | Relative percentual error – cluster validations

Cluster	Q_{100}	Q_{90}	Q_{80}	Q_{70}	Q_{60}	Q_{50}	Q_{40}	Q_{35}	Q_{30}	Q_{20}	Q_{10}	Q_5	Q_1	$Q_{0.5}$	$Q_{0.1}$
1	36.1	86.5	57.4	48.8	44.5	40.6	27.9	12.2	10.3	25.5	16.9	2.0	39.5	54.4	112.1
2	143.0	0.4	1.9	3.5	0.1	2.1	2.9	2.2	5.8	4.0	1.2	2.9	7.1	7.3	11.7
3	368.4	20.2	0.3	17.1	12.7	13.7	17.4	13.8	13.2	9.6	17.1	9.6	30.7	34.2	13.9
4	99.7	71.4	62.9	65.0	60.7	29.5	16.7	11.6	17.2	39.7	15.5	16.9	43.4	44.2	57.2
5	36.5	10.0	9.7	14.9	15.7	11.7	5.9	2.9	8.3	3.5	13.3	17.9	47.0	45.4	55.7
6	70.7	18.9	18.5	6.1	3.3	4.5	5.8	3.0	7.0	11.2	5.4	0.5	18.6	18.5	34.1
7	33.0	67.5	62.7	60.0	57.3	19.6	4.4	24.4	28.2	30.8	22.2	37.6	128.2	130.7	162.1
8	123.8	6.7	0.8	1.6	8.1	9.3	8.2	9.5	4.8	1.6	4.6	13.3	7.5	20.2	22.2
9	188.0	67.7	66.5	25.6	10.8	0.8	3.3	4.6	4.4	7.3	8.3	9.4	0.3	0.5	12.0
10	96.6	68.8	22.3	15.5	21.9	30.4	37.3	32.7	32.6	9.6	2.1	49.0	66.1	12.8	1.0
11	43.6	18.1	23.5	29.4	20.5	2.6	2.3	7.7	8.7	8.4	15.6	8.6	22.4	2.0	0.4
12	59.3	34.8	0.6	6.0	0.5	9.2	6.8	4.9	4.4	3.7	8.1	7.3	12.9	9.1	12.2
13	85.1	17.0	21.5	8.1	1.9	4.0	0.6	2.7	4.4	5.3	10.8	5.7	0.1	2.6	22.7
14	76.4	4.3	8.7	4.7	22.5	10.2	3.9	2.1	0.4	7.6	4.4	100.0	47.9	37.5	66.6
15	1475.6	157.1	81.0	59.2	44.7	34.3	6.4	8.6	8.3	5.8	19.4	27.9	35.9	41.9	54.1
Average	*195.7*	*43.3*	*29.2*	*24.3*	*21.7*	*14.8*	*10.0*	*9.5*	*10.5*	*11.6*	*11.0*	*20.6*	*33.8*	*30.8*	*42.5*

Please refer to the online version of this paper to see this table in colour: http://dx.doi.org/10.2166/nh.2022.022.

Soil use and land cover variables are notably frequent in regression equations, followed by climatic variables from precipitation and evapotranspiration, then by topographic variables like elevations and slope. Regression equations are helpful for understanding which variables and processes involved are more relevant in the flow regime and basin's rainfall–runoff estimations.

Better results were obtained for mean streamflow percentiles compared with maximum and minimum flows. A possible reason for this is that the gauging stations were calibrated frequently in average flows and not in flood events ($Q_{p<1}$) or in recessions ($Q_{p>85}$). However, these are extrapolations in observed FDC. Nevertheless, variations in mean R^2 values among flow magnitudes were insignificant, excepting a lower value in percentile 20 (Figure 3), which can be defined as a transition between the average and maximum discharges.

The form of dimensionless FDC was strongly related to basin size. It can be observed that small basins (magnitude order between 10^1 and 10^2 km^2) FDC had an 'L' form (pronounced concavity), which represents a high difference between extreme and mean flows, whereas big basins (10^4 and 10^5 km^2) FDC form was softened. The FDC form resulted from the geomorphological attributes and their interrelations (Perez *et al.* 2018), which are very complex.

Table 9 | Relative percentual error – traditional subregions validations

Region	Q_{100}	Q_{90}	Q_{80}	Q_{70}	Q_{60}	Q_{50}	Q_{40}	Q_{35}	Q_{30}	Q_{20}	Q_{10}	Q_5	Q_1	$Q_{0.5}$	$Q_{0.1}$
Caribe	17.86	29.6	13	5.4	4.1	7.7	9	6.6	21	13	2.5	9.18	19.6	3.94	7.9
Magdalena	64.41	16.9	11	11	5.4	1.7	0.3	0.3	8.9	1.2	4.5	4.25	15.9	14.9	41.3
Orinoquia	34.44	13.4	34	33	25	17	12	3.8	1.1	3.1	9.9	15.5	15.5	19.4	1.7
Amazonia	59.35	35.3	14	16	7.9	7.4	4.6	2.3	11	3.2	16	7.39	43.4	54.8	69.5
Pacific	417.5	15.1	12	8.8	7	6.3	4.5	4.2	3.8	2.7	8.4	14.1	16.6	15	10.9
Average:	*118.7*	*22.1*	*17*	*15*	*9.9*	*8*	*6*	*3.4*	*9.1*	*4.7*	*8.3*	*10.1*	*22.2*	*21.6*	*26.2*

Please refer to the online version of this paper to see this table in colour: http://dx.doi.org/10.2166/nh.2022.022.

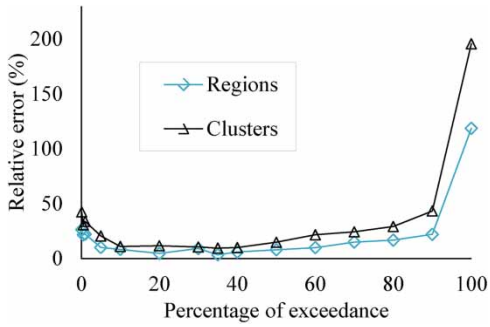

Figure 11 | Comparison between relative error in cluster and traditional grouping estimation.

Some estimations show a Q_{p1} with a higher percentage of exceedance than Q_{p2} (p1>p2). Here a conceptual error is induced. If something like this occurs in FDC estimation, the wrong percentile should be discarded, and linearly interpolated between neighbors should be used. Errors like this were uncommon in this work's validations.

The correlations obtained between observed and estimated flows and $y = x$ matrices were high for clusters and regions (Tables 6 and 7). These high values indicate coherence in magnitude order in dimensionless flow estimations.

The regression model estimates dimensionless flow regime; however, to get the original FDC (flow in m^3/s), the product of each Q^* must be calculated using the long-term mean flow. There is a possibility that the approximate long-term average flow of each Colombian basin can be determined by applying a long-term water balance. However, estimating mean annual precipitation and real evapotranspiration represents an additional error source.

5. CONCLUSIONS

Despite Colombia's wide lack of hydrological information, it is possible to extrapolate conditions to perform flow regime estimation in ungauged sites. It is assumed that selected gauging stations and their respective hydrological basins span into a large spectrum of characteristics like size, form, mean discharges, and climatic conditions, which allows us to conduct the approach with similar success probabilities for different characteristics of rivers.

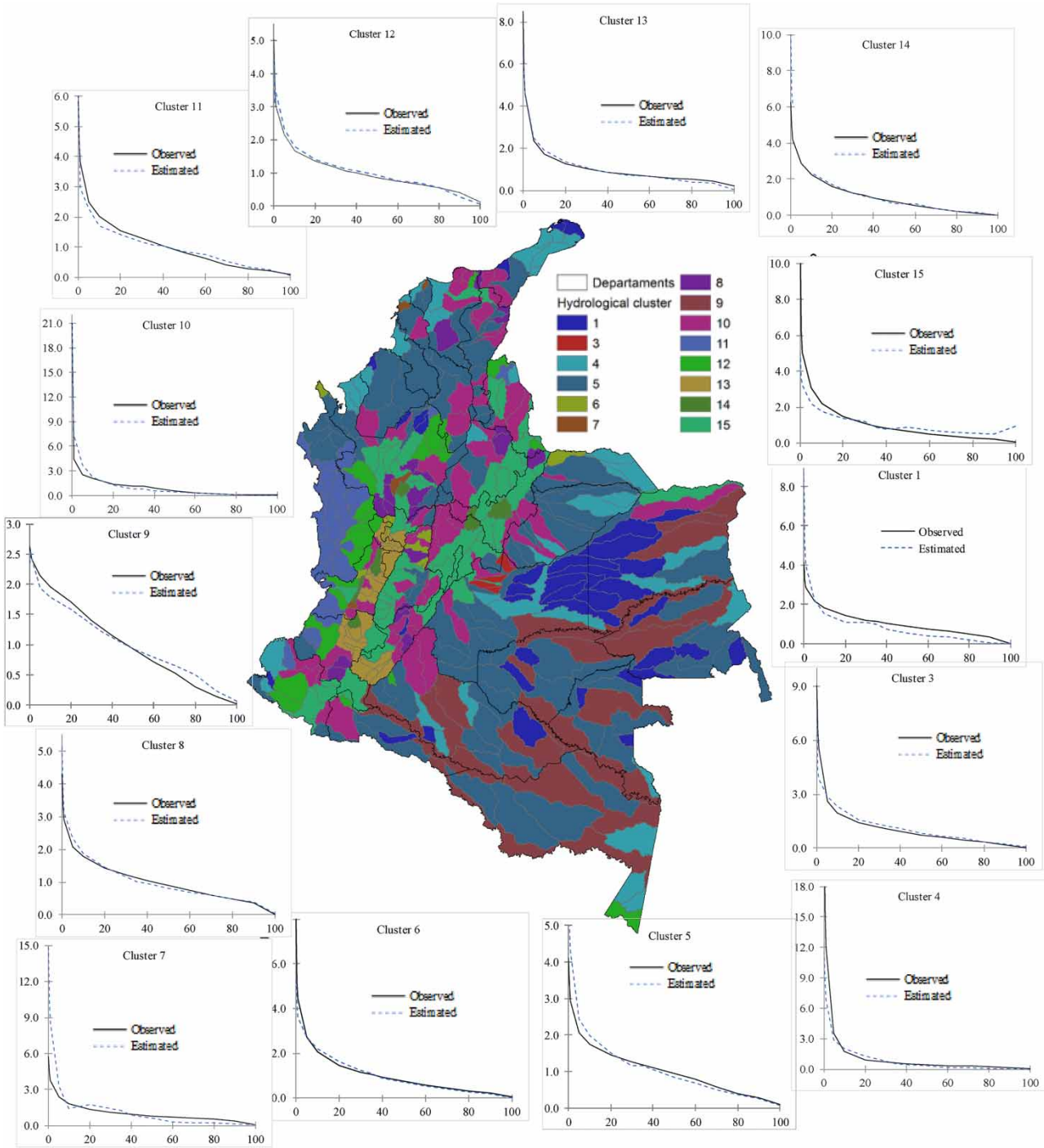

Figure 12 | FDC estimation results summary map. Please refer to the online version of this paper to see this table in colour: http://dx.doi.org/10.2166/nh.2022.022.

An equation series that allows FDC estimation in ungauged catchments in Colombia was performed. Attributes used in this work are publicly available in web databases. This methodology requires defining a targeted basin, locating it on a cluster using geometric centroid coordinates, searching for the estimation equations, and defining which attributes are needed. Attribute units and dimensions are specified.

Traditionally, five homogeneous hydrological regions are defined in Colombian territory. Nevertheless, too many variables are involved and related to the basin's behavior, which has a heterogeneous spatial distribution. The closeness criterion was not considered in the homogeneous hydrologic regions.

The way MLR was applied in this work represents an adjustment in the Colombian flow regime sectioned estimations. There are other options for this, for example, spline regressions. According to model validation results, low mean percentage errors were achieved for average discharge percentiles. Nevertheless, extreme values (minimum and maximum flow) gave higher mean error values because their flows were not measured but extrapolated.

Basin size is influential in flow regime estimation, even in dimensionless flow. Aggrupation and estimation models consider geographic basin extension, and parameter drainage area is not related only to basin size. For example, basin perimeter, Graveluis compactness coefficient, mainstream length, tectonic fault density, and basin unevenness are closely related too. There is an area dependency in these attributes, which is observed in the results of α and β values. May review in the Supplementary Material on Gaviria Arbeláez (2019).

ACKNOWLEDGEMENTS

Analyses are based on data provided by IDEAM, SGC, and IGAC national institutes. We are beforehand grateful to potential reviewers and editors.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Álvarez, G., Hotait, N. & Sustaita, F. 2011 Identificación de regiones hidrológicas homogéneas mediante análisis multivariado. *Ingeniería Investigación y Tecnología* **XII** (3), 277–284.

Anderson, T. 1958 *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, Inc, New York, London, Sydney.

Barco, O. J. & Cuartas, L. A. 1998 Estimación de la Evaporación en Colombia. Trabajo Dirigido de Grado. Universidad Nacional de Colombia, Sede Medellín.

Beven, K. J. 2012 *Rainfall-Runoff Modelling : The Primer*. Wiley-Blackwell, Oxford.

Blöschl, G. 2005 Rainfall-runoff modeling of ungauged catchments. *Encyclopedia of Hydrological Sciences*. https://doi.org/10.1002/0470848944.hsa140.

Burt, T. P. & Swank, W. T. 1992 Flow frequency responses to hardwood-to-grass conversion and subsequent succession. *Hydrological Processes* **6** (2), 179–188. https://doi.org/10.1002/hyp.3360060206.

Castellarin, A., Camorani, G. & Brath, A. 2007 Predicting annual and long-term flow-duration curves in ungauged basins. *Advances in Water Resources* **30** (4), 937–953. https://doi.org/10.1016/J.ADVWATRES.2006.08.006.

Castellarin, A. 2014 Regional prediction of flow-duration curves using a three-dimensional kriging. *Journal of Hydrology* **513**, 179–191. https://doi.org/10.1016/J.JHYDROL.2014.03.050.

Chaves, C. B. & Jaramillo, R. A. 1998 Regionalización de la temperatura del aire en Colombia. *Repositorio Digital Del Centro Nacional de Investigación Del Café – Cenicafé* **49** (3), 224–230.

Doulatyari, B., Betterle, A., Basso, S., Biswal, B., Schirmer, M. & Botter, G. 2015 Predicting streamflow distributions and flow duration curves from landscape and climate. *Advances in Water Resources* **83**, 285–298. https://doi.org/10.1016/j.advwatres.2015.06.013.

Foster, H. A. 1933 Duration curves. *Proceedings of the American Society of Civil Engineers* **59** (8), 1223–1246.

García, P. L., Méndez, J. F. & Zárate, M. F. 2017 Delimitation of Colombia hydrologic regions. *Ingeniería Y Desarrollo* **35** (1), 132–151. https://doi.org/10.14482/inde.35.1.8946.

Gaviria Arbeláez, C. J. 2019 *Regionalización de Curvas de Duración de Caudales en Colombia*. Universidad Nacional de Colombia sede Medellín. Available from: http://bdigital.unal.edu.co/71551/1/1039458525.2019.pdf

GWP. 2008 *Principios de gestión integrada de los recursos hídricos. Bases para el desarrollo de planes nacionales*.

Hartigan, J. 1975 *Clustering Algorithms*. Available from: http://cds.cern.ch/record/105051

Hurtado, A. F. & Mesa, Ó. J. 2014 Reanalysis of monthly precipitation fields in Colombian territory. *DYNA* **81** (186), 251. https://doi.org/10.15446/dyna.v81n186.40419.

IDEAM 2008 *Informe Anual sobre el Estado del Medio Ambiente y los Recursos Naturales Renovables en Colombia, Estudio nacional del agua: Relaciones de demanda de agua y de oferta hídrica*.

IDEAM 2014 *Estudio Nacional del Agua 2014*.

IDEAM, & MinAmbiente 2015 Análisis integrado. In: Ministerio de Ambiente, Bogotá (ed.). Instituto de Hidrologia, Meteorología y Estudios Ambientales *Estudio Nacional del Agua 2014*. Instituto de Hidrologia, Meteorología y Estudios Ambientales.

IGAC. 2014 *Datos Abiertos Cartografía y Geografía. Cartografía Básica de Colombia escala 1:100.000*.

Jaramillo, A. 1989 Relación entre la evapotranspiración y los elementos climáticos. (Nota técnica). *Cenicafé* **40** (3), 288–298.

Li, M., Shao, Q. & Zhang, L. 2010 A new regionalization approach and its application to predict flow duration curve in ungauged basins. *Journal of Hydrology*. Elsevier **389** (1–2), 137–145.

Mesa Sánchez, Ó. J., Vélez Upegui, J. I., Giraldo Osorio, J. D. & Quevedo Tejada, D. I. 2003 *Regionalización de Características Medias de la Cuenca con Aplicación a Estimación de Caudales Máximos*. Repositorio Institucional Universidad Nacional de Colombia, Medellín.

Mohamoud, Y. M. 2008 Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves. *Hydrological Sciences Journal* **53** (4), 706–724. https://doi.org/10.1623/hysj.53.4.706.

Musiake, K., Inokuti, S. & Talahasi, Y. 1975 Dependence of low flow characteristics on basin geology in mountainous areas of Japan. In Proceedings of International Symposium of Hydrology, Tokyo, Japan, 1975, 117, pp. 147–156.

NASA & Watkins, D. 2014 *Datos SRTM, resolución 30 y 90 metros*.

Olden, J. D., Kennard, M. J. & Pusey, B. J. 2012 A framework for hydrologic classification with a review of methodologies and applications in ecohydrology. *Ecohydrology* **5** (4), 503–518. https://doi.org/10.1002/eco.251.

Perez, G., Mantilla, R. & Krajewski, W. F. 2018 The influence of spatial variability of width functions on regional peak flow regressions. *Water Resources Research* **54** (10), 7651–7669. https://doi.org/10.1029/2018WR023509.

Perez, G., Mantilla, R., Krajewski, W. F. & Quintero, F. 2019 Examining observed rainfall, soil moisture, and river network variabilities on peak flow scaling of rainfall-runoff events with implications on regionalization of peak flow quantiles. *Water Resources Research* **55** (12), 10707–10726. https://doi.org/10.1029/2019WR026028.

Razavi, T. & Coulibaly, P. 2013 Streamflow prediction in ungauged basins: review of regionalization methods. *Journal of Hydrologic Engineering* **18** (8), 958–975. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690.

Rojo Abuín, J. M. 2007 Regresión Lineal Múltiple. *Institto de Economía Y Gografía* **2**, 25.

Salazar-Holguín, F. 2013 *Zonificación Hidrográfica Preliminar de Colombia*. IDEM, Bogotá.

Salazar Oliveros, J. 2016 Una metodología para la estimación de curvas de duración de caudales (cdc) en cuencas no instrumentadas. Caso de aplicación para Colombia en los departamentos de Santander y Norte de Santander. *Repositorio Institucional – Universidad Nacional de Colombia*.

Sauquet, E. & Catalogne, C. 2011 Comparison of catchment grouping methods for flow duration curve estimation at ungauged sites in France. *Hydrology and Earth System Sciences* **15** (8), 2421–2435.

Sivapalan, M. 2003 Prediction in ungauged basins: a grand challenge for theoretical hydrology. *Hydrological Processes* **17** (15), 3163–3170. https://doi.org/10.1002/hyp.5155.

Steel, R. G. D. & Torrie, J. H. 1960 *Principles and Procedures of Statistics*. McGRAW-Hill Book Company, Inc, New York, Toronto, London.

Swain, J. B. & Patra, K. C. 2017 Streamflow estimation in ungauged catchments using regional flow duration curve: comparative study. *Journal of Hydrologic Engineering* **22** (7), 04017010. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001509.

UPME-PPUJ 2015 *Atlas potencial hidroenergético de Colombia*. Colciencias, Bogotá.

Vogel, R. M. & Fennessey, N. M. 1994 Flow-duration curves. I: new interpretation and confidence intervals. *Journal of Water Resources Planning and Management* **120** (4), 485–504. https://doi.org/10.1061/(ASCE)0733-9496(1994)120:4(485).

Wagener, T., Blöschl, G. & Sivapalan, M. 2013 *Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales*. Cambridge Univ. Press, NY, pp. 11–28.

Wilks, D. S. 2011 Cluster analysis. *International Geophysics* **100**, 603–616. https://doi.org/10.1016/B978-0-12-385022-5.00015-4.

Winter, T. C. 2001 The concept of hydrologic landscapes. *Journal of the American Water Resources Association* **37** (2), 335–349. https://doi.org/10.1111/j.1752-1688.2001.tb00973.x.

Wood, E. F., Sivapalan, M. & Beven, K. 1990 Similarity and scale in catchment storm response. *Reviews of Geophysics* **28** (1), 1. https://doi.org/10.1029/RG028i001p00001.

First received 10 February 2022; accepted in revised form 23 May 2022. Available online 30 July 2022