A fuzzy hybrid clustering method for identifying hydrologic homogeneous regions

S. Saeid Mousavi Nadoushani, Naser Dehghanian and Bahram Saghafian

ABSTRACT

Identification of hydrologic homogeneous regions (HHR) facilitates prioritization of watershed management measures. In this study, a new methodology involving a combination of self-organizing features maps (SOFM) method and fuzzy C-means algorithm (FCM), designated as SOMFCM, is presented to identify HHRs. The case study region is Walnut Gulch Experimental Watershed (WGEW) located in Arizona. The input data consisted of a number of factors that influence runoff generation processes, including ten surface features as well as various rainfall values corresponding to 25, 50, and 100 years return periods. Factor analysis (FA) was applied for the selection of effective surface features along with rainfall value, used in the clustering algorithm. Validation procedure indicated that the best clustering scenario was achieved through merging three layers including TPI (topographic position index), CN (curve number), and P50 (50-year rainfall). The optimum number of clusters turned out to be six while the fuzzification parameter became 1.6. The presented methodology may be proposed as a simple approach for identifying HHRs.

Key words | factor analysis (FA), fuzzy C-means (FCM), hydrologic homogeneous regions (HHR), rainfall-runoff process, self-organizing features maps (SOFM)

S. Saeid Mousavi Nadoushani (corresponding author)
Naser Dehghanian
Faculty of Water and Environmental Engineering,
Shahid Beheshti University,
Tehran,
Iran
E-mail: *sa_mousavi@sbu.ac.ir*

Bahram Saghafian
Department of Technical and Engineering, Science and Research Branch,
Islamic Azad University,
Tehran,
Iran

ACRONYMS AND ABBREVIATIONS

ANN	Artificial Neural Networks
CN	Curve Number
DEM	Digital Elevation Model
FA	Factor Analysis
FCM	Fuzzy C-means Algorithm
HHR	Hydrologic Homogeneous Regions
HLR	Hydrologic Landscape Regions
HRU	Hydrological Response Units
KMO	Kaiser–Meyer–Olkin statistic
Ksat	Soil saturated hydraulic conductivity
MARD	Mean Absolute Relative Distance
NDVI	Normalized Difference Vegetation Index
P25	The 24-hour rainfall with return period of 25 years
P50	The 24-hour rainfall with return period of 50 years
P100	The 24-hour rainfall with return period of 100 years
PCA	Principal Component Analysis
REA	Representative Elementary Area
REW	Representative Elementary Watersheds
SOFM	Self-Organizing Features Maps
TPI	Topographic Position Index
VAR	Variable
WGEW	Walnut Gulch Experimental Watershed

doi: 10.2166/hydro.2018.004

INTRODUCTION

In recent decades, water resources studies at watershed scale have been enhanced through the deployment of geographic information systems, satellite imagery data, and water-related softwares. In this regard, spatially distributed

analysis and tools are becoming vital for hydrologists and watershed managers. Accordingly, classification and clustering of the watershed landscape into homogeneous regions facilitates prioritization of watershed management measures.

In recent decades, hydrological spatial discretization has also enjoyed new terminologies such as representative elementary area (REA) by Wood *et al.* (1988), hydrological response units (HRU) by Flügel (1995), representative elementary watersheds (REWs) by Reggiani *et al.* (1998, 1999, 2000), hydrologic landscape regions (HLRs) by Wolock *et al.* (2004), REWs for cold and snowy regions by Tian *et al.* (2006), a flexible methodology for discretization into REWs by Dehotin & Braud (2008) and Kuentz *et al.* (2017). In the meantime, clustering algorithms such as Ward algorithm (Ward 1963), self-organizing feature mapping (SOFM) method (Kohonen 1982), and fuzzy c-means (FCM) clustering algorithm (Bezdek 1981) have been used in numerous applications.

The FCM method has been widely adopted in natural sciences (e.g., Bruin & Stein 1998), discretization of natural watersheds into groups (e.g., Rao & Srinivas 2006), water resources (e.g., Golian *et al.* 2010; Irwin 2015), and regionalization studies (e.g., Basu & Srinivas 2015). One of the advantages of FCM is the ability to determine the cells/clusters uncertainty through confusion index. However, the main FCM weakness is the way to determine the FCM's parameters, i.e., the fuzzification parameter and the optimum number of clusters. Therefore, fuzzy clustering, similar to other clustering methods, may not be used without determination of the optimum number of clusters. Fuzzy clustering, FCM in particular, has been recommended to partition watersheds into natural landscapes or groups with the homogeneous hydrologic response by different researchers (e.g., Bruin & Stein 1998; Rao & Srinivas 2008).

The SOFM has also been used in numerous fields, including in water resources applications (e.g., Nourani & Parhizkar 2013; Nourani *et al.* 2016; Sharghi *et al.* 2018). This method is considered as a useful way to determine homogeneous regions as it enjoys high performance in visualizing and summarizing the contributing features as well as the ability to display the distribution of each component (Farsadnia *et al.* 2014).

Hybrid clustering methods, for example, the combination of SOFM and FCM, have been successfully used in the regionalization of sites or catchments (e.g., Srinivas *et al.* 2008; Farsadnia *et al.* 2014; Ahani & Nadoushani 2016). The SOFM-FCM combination can identify homogeneous regions more quickly and accurately while having sufficient flexibility to minimize the FCM objective function.

The main objective of this study is to derive hydrologic homogeneous regions (HHRs) through the adoption of effective rainfall-runoff characteristics. This is achieved via application of a fuzzy hybrid clustering method that will be evaluated through new criteria. In this regard, in addition to the application of well-known and conventional clustering validation approaches traditionally developed and used for FCM, two and one technique(s) are proposed for determining the optimum number of clusters and the fuzzification parameter, respectively. Furthermore, with emphasis on the outcome of this study, the advantages of integrating FCM with the SOFM are outlined.

METHODOLOGY

Factor analysis

Factor analysis (FA) is one of the early multivariable statistical approaches that are able to determine the data structure by summarizing and reducing the data size (Suhr 2006). This method is frequently applied in climatic (e.g., Bartzokas *et al.* 2003) and hydrologic studies (e.g., Farsadnia *et al.* 2014). FA method is fully described by Harman (1976) and Rencher (1995).

FA can improve the effectiveness of SOFM classification, leading to an increase in accuracy of cluster centers. In this study, clustering is performed based on the most effective variables resulting from the FA pre-processing.

FCM

Natural geographical landscapes, unlike the man-made domains, often do not have a clear border. Therefore, the use of fuzzy clustering may be more justified than hard clustering (crisp) in the discretization of terrestrial and, in particular, hydrological regions.

Among fuzzy clustering techniques, FCM algorithm, proposed by Bezdek (1981), is simple and comprehensive and is mostly used on uncertain data sets. In this study, the Euclidean distance criterion, deemed more appropriate to cluster the data, with no linear relationship is adopted. Since the input data units vary in clustering with distance criterion, inputs with larger numbers will be dominant, producing a deviation in results. Therefore, standardization or rescaling is performed as follows:

$$x_{jn} = \frac{y_{jn} - y_{n.min}}{y_{n.max} - y_{n.min}} \tag{1}$$

where x_{jn} is the rescaled value of y_{jn} (actual value), y_{jn} is the nth data of the jth input layer, and $y_{n,min}$ and $y_{n,max}$ are the minimum and maximum actual values of the jth input layer, respectively. In FCM algorithm, the aim is to minimize the following objective function (OF):

$$J = \sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij}^{m} d_{ij}^{2} = \sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij}^{m} \|\boldsymbol{x_j} - \boldsymbol{c_i}\|^2 \tag{2}$$

which is subject to the following constraints:

$$J = \sum_{i=1}^{c} u_{ij} = 1 \ \forall\ j \in \{1\ .\ \ldots .\ N\} \tag{3}$$

$$0 < \sum_{j=1}^{N} u_{ij} < N \ \forall\ i \in \{1\ .\ \ldots .\ c\} \tag{4}$$

where N is the number of cells in each layer, c is the number of clusters, $u_{ij} \in [0.1]$ is the degree of membership dedicated to jth rescaled feature cell belonging to the ith cluster represented by its center c_i, m is the weight exponent or fuzzification parameter for each fuzzy membership which determines the fuzziness of the clusters and is a real value between [1, ∞). The value of $m = 2$ has been widely chosen as a default value for fuzzification parameter (Hathaway & Bezdek 2001). However, there is no theoretical basis for the optimum selection of m. Also, x_j is the value of rescaled feature cell in the jth layer, c_i is the center of the ith cluster, and d_{ij} or $|x_j - c_i|$ is the Euclidean distance which is the distance between cells and values corresponding to the cluster centers.

Forcing the derivative of the OF (Equation (2)) to zero, one gets:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{m-1}}} \ for\ i = 1.\ \ldots .\ c\ \&\ k = 1\ .\ \ldots .\ N \tag{5}$$

$$c_i = \frac{\sum_{j=1}^{N} u_{ij}^{m} x_j}{\sum_{j=1}^{N} u_{ij}^{m}} \ for\ i = 1.\ \ldots .\ c \tag{6}$$

FCM is performed through an iterative procedure of optimizing the OF by updating the degree of membership u_{ij} (Equation (5)) and the cluster centers c_i (Equation (6)). The procedure has been described in detail by Bezdek (1981).

In order to ensure proper clustering, special attention needs to be paid to: (1) determine the fuzzification parameter (m); (2) select the degree of membership for the winning cluster (degree of membership threshold); (3) control clustering uncertainty; (4) validate fuzzy clustering to optimize the number of clusters; and ultimately (5) measure the similarity of cells to its cluster center as a second control.

Chiu (2005) proposed two α-cut (α_{low}, α_{high}) for the degree of membership based on the number of clusters (c). If $\alpha_{high} \geq 1-1/c$, then that cell definitely belongs to the cluster, and if $\alpha_{low} < 1/c$, the membership of a cell to the cluster is decisively rejected. Therefore, the value of $1/c$ is believed to be an acceptable choice for the threshold fuzzy membership (Rao & Srinivas 2008). In this study, the cells which firmly belong to each cluster are determined by specifying a threshold fuzzy membership value equal to $1/c$ ($c = 4, \ldots, 11$). Therefore, the cells in the cluster which have the memberships greater than the specified threshold value can produce a fuzzy cluster.

In order to analyze the uncertainty of fuzzy clustering, the following indicator, known as the confusion index (CI) (Burrough et al. 1997) may be used:

$$CI_i = 1 - (u_{i.1st\ max} - u_{i.2nd\ max})\ .\ CI \in [0.1] \tag{7}$$

where $u_{i,1st\ max}$ and $u_{i,2nd\ max}$ are the first and second highest degree of membership values for cell i, respectively. High values of CI (close to 1) represent greater confusion. CI represents the degree of certainty of a cell to belong to a cluster,

and also indicates the overlapping of adjacent clusters in the cluster boundaries.

For qualitative evaluation of this parameter, the *CI* values in the interval [0,1] may be divided into five classes denoting insignificant (0–0.2), very low (0.2–0.4), low (0.4–0.6), high (0.6–0.8), and very high (0.8–1) uncertainty.

Fuzzy clustering validation indices

Validation in fuzzy clustering involves maximization of two important criteria: first, the density of the data in each cluster and second, the distance between the clusters, which indirectly leads to the optimization of the number of clusters. Various fuzzy cluster validation indices have been proposed to determine optimum c in a data set (Pal & Bezdek 1995). In the present study, five cluster validity indices, as presented in Table 1, are examined.

SOFM

Nonlinearity and flexibility of artificial neural networks (ANN) are attractive features in a variety of applications in hydrology (Govindaraju & Rao 2000). A particular category of ANN, known as SOFM (Kohonen 1982), has been used as a clustering tool in order to convert the complex nonlinear statistical relationship among high-dimensional input data into a simple and geometric relationship on a low-dimensional display (Nourani *et al.* 2015).

SOFMs are one of the unsupervised training networks such that no output is required to classify the existing data, yet they convert complex multi-dimensional data into visible clusters. SOFMs attempt to find the topological structure in the input data by mapping the available data on an attribute mapping, which may also be considered as the output layer (Rao & Srinivas 2008).

In the absence of clear understandable patterns in existing data, the interpretation of SOFM clusters, regardless of size and dimensions, is rarely possible. However, in such cases, SOFMs may be considered as a useful tool for supervised clustering algorithms (Rao & Srinivas 2008).

In this study, SOFM method was used to provide the initial cluster centers for FCM algorithm. The combination of SOFM and FCM is denoted as SOMFCM, that was assembled using R features including 'fclust' (Ferraro & Giordani 2015), 'kohonen' (Wehrens & Buydens 2007), 'cluster' (Maechler *et al.* 2015), and 'e1071' (Meyer *et al.* 2015).

Table 1 | Selected fuzzy clustering validation indices for optimizing the number of clusters (c)

Index	Source	Equation		Parameters
Partition coefficient	Bezdek (1974a, 1974b)	$V_{PC} = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{c} u_{ij}^2$	(8)	V_{PC}: Partition coefficient
Partition entropy	Bezdek (1974a, 1974b)	$V_{PE} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{c} u_{ij}\cdot\log(u_{ij})$	(9)	V_{PE}: Partition entropy
Fukuyama–Sugeno	Fukuyama & Sugeno (1989)	$V_{FS} = \sum_{j=1}^{N}\sum_{i=1}^{c} u_{ij}^m(\|x_j - c_i\|^2 - \|c_i - \bar{c}\|^2)$ $\bar{c} = \sum_{j=1}^{N}\frac{x_j}{N}$	(10)	V_{FS}: Fukuyama and Sugeno validity measure V_{EXB}: The extended FCM
Extended Xie–Beni	Xie & Beni (1991) and Pal & Bezdek (1995)	$V_{EXB} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij}^m\|x_j - c_i\|^2}{N\min_{i\neq j}\|c_j - c_i\|^2}$	(11)	Xie–Beni index V_K: Kwon index N: Number of cells in each layer
Kwon's index	Kwon (1998)	$V_k = \frac{\sum_{j=1}^{N}\sum_{i=1}^{c} u_{ij}^2\|c_i - x_j\|^2 + \frac{1}{c}\sum_{i=1}^{c}\|c_i - \bar{c}\|^2}{\min_{i\neq j}\|c_i - c_j\|^2}$ $\bar{c} = \sum_{j=1}^{N}\frac{x_j}{N}$	(12)	c: Number of clusters u_{ij}: Degree of membership x_j: Rescaled feature value for each cell c_i and c_j : The fuzzy centroids of the ith cluster

SOMFCM validation

Generally speaking, determination of the number of clusters before clustering is a disadvantage of most clustering techniques (Irwin 2015). Furthermore, the main disadvantage of FCM algorithm is its sensitivity to the fuzzification parameter (m) and the number of clusters (c), both of which should be optimized simultaneously. As m increases, the degree of membership of a given cell to a cluster decreases, causing overlap of the membership of the cell to adjacent clusters in boundary regions (the overlapping of adjacent clusters leads to the gradual display of variations in nature). In contrast, as c increases, the similarity of cells within a cluster to its cluster center increases and the entropy in each cluster decreases, while the average cell values of the membership degree also decrease.

Accordingly, the following SOMFCM validation techniques are adopted in order to determine m and c.

Technique I (optimum m): As fuzzification parameter (m) increases, the V_{PC} (Equation (8))/V_{PE} (Equation (9)) decreases/increases, so that the intersection point of simultaneous changes of the two for different values of m and c may be identified as the optimum m. Rao & Srinivas (2008) also suggested the indirect use of V_{PC} and V_{PE} in order to determine the range of m and its optimum value in a fuzzy clustering context.

Technique II (optimum c): If the number of clusters (c) increases, the distance between the scaled values of the cells and that of the cluster center (expressed as the mean Euclidean distance, *Euc_mean*) decreases, whereas the mean confusion index (*CI_mean*) of the cells increases. Therefore, the interaction of these two (namely, *CI_mean* and *Euc_mean*) for different values of c and m may yield the optimum c.

Technique III (optimum c): The correlation coefficient R^2, defined as the ratio of the variance of the data corresponding to the cluster centers to the variance of the actual data, may be considered as one criterion in determining the optimum c:

$$R^2 = \frac{\sum (c_i - \bar{x}_j)^2}{\sum (x_j - \bar{x}_j)^2} \tag{13}$$

The third technique involves interaction between the number of cells with the degree of membership equal to or greater than α_{high} (in this study $\alpha_{high} \geq 1-1/c$) expressed by *N_Umax* with the value of R^2 that can be simultaneously drawn for different values of c and m. When c is reduced, the *N_Umax* typically decreases while R^2 increases.

Study area

The Walnut Gulch Experimental Watershed (WGEW) was selected as a research area by the United States Department of Agriculture (USDA) in the mid-1950s. The WGEW with an area of 150 square kilometers is located in southeastern Arizona, USA, at 31°42′ N latitude and 110°03′ W longitude.

The elevation of the watershed ranges between 1,221 m and 1,887 m. The climate is classified as semi-arid, with a mean annual temperature of 17.7 °C and mean annual precipitation of 350 mm. The precipitation regime is dominated by the North American Monsoon with slightly more than 60% of the annual total occurring during July–September (USDA 2007).

The available literature reports on various aspects of the watershed including Goodrich *et al.* (2008) on precipitation, Stone *et al.* (2008) on runoff, Skirvin *et al.* (2008) on vegetation, Keefer *et al.* (2008) on meteorology and soil, and Heilman *et al.* (2008) on GIS maps. The quality of watershed data has been assured and that was the motivation for choosing this watershed in the current study.

Hydrological, meteorological, soil and land use data are available online at http://www.tucson.ars.ag.gov/dap. Figure 1 shows the digital elevation model (DEM), along with the distribution of 26 selected rain gauges within or close to WGEW.

Data sets

The selection of input data required for discretization of HHRs depends on the studied hydrological process. In this study, rainfall-runoff processes are the main focus, thus effective layers must deal with meteorological attributes as well as watershed physiographic/surface features. The scale of the maps was 1:24,000 and the selected cell size was 150 meters for all watershed input layers. The

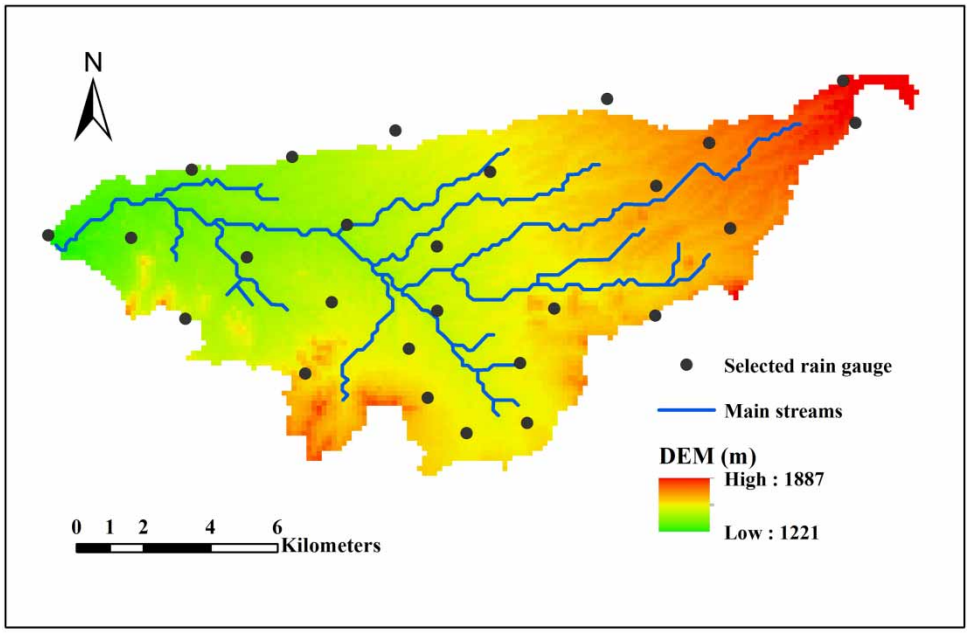

Figure 1 | DEM and selected rain gauges (cell size is 150 m).

characteristics of the input layers entered in factor analysis are shown in Table 2.

Procedure

The flow chart of the study procedure is presented in Figure 2.

RESULTS AND DISCUSSION

Results are presented in three sections: input data preprocessing, factor analysis, and clustering.

Input data preprocessing

In the present study, the selected cell size was 150 (m) for all input layers. This selection was based on the recommendation of Bloschl & Sivapalan (1995) in order to consider compatibility between the scale of observed data with scale and purpose of the simulation of hydrological processes. On the other hand, as hilly lands cover most parts of the WGEW and the river network is relatively dense, use of cell sizes larger than 150 m may introduce uncertainty in the spatial aggregation. Some selected input features are shown in Figure 3.

Since rainfall is the most important factor in rainfall-runoff events, it must be taken into account in the prioritization of the regions in terms of potential runoff generation (Saghafian & Khosroshahi 2005). In this study, rainfall layers corresponding to 25-, 50- and 100-year return periods were generated using rainfall frequency analysis of available 50-year rainfall time series (1967–2016) for 26 selected stations inside and outside the watershed (as shown in Figure 1).

According to the results of rainfall frequency analysis, the fitted distribution functions for 26 rain gauges were Log Pearson 3 (14 stations), Lognormal (5 stations), Gamma (5 stations), and Gumbel (2 stations). The 24-hour rainfall with return periods of 25-, 50- and 100-year was estimated for each rain gauge. Figure 4, for example, shows the 50-year rainfall map.

Factor analysis

The total number of cells entering factor analysis was equal to 6,225 for each component/variable. Then, each of the ten attributes was rescaled by Equation (1), thus forming a $6{,}225 \times 10$ matrix. To determine whether factor analysis was an appropriate tool for data reduction, Kaiser–Meyer–Olkin (KMO) statistic of sampling adequacy was computed. The KMO statistic equaled 0.922, that confirmed the applicability

Table 2 | Input data related to surface characteristics of the WGEW study watershed

No.	Input data	Unit	Description
1	DEM	m	Digital elevation model (DEM) with 150 m cell size
2	Normalized difference vegetation index (*NDVI*)	–	Constructed based on LANDSAT bands 3 and 4 (red and near-infrared) corresponding to the time of historic floods (July and August 2007)
3	Soil saturated hydraulic conductivity (*Ksat*)	mm/hr	Available on WGEW-USDA website prepared in four categories with a range of (8.36)
4	Topographic position index (*TPI*)	m	*TPI* reflects the tendency of each cell to saturate. According to Jenness (2006), the *TPI*, initially developed by Weiss (2001), enjoys simplicity in form and application yet is very powerful in categorizing the morphometry of natural regions and useful for runoff and erosion studies
5	Slope aspect	–	In 9 categories (1 flat area and 8 other main directions)
6	Plan curvature	–	Influences the convergence/divergence of surface runoff
7	Profile curvature	–	Influences the runoff acceleration/deceleration
8	Slope	Percent	Slope (in percent)
9	Curve number (*CN*)	–	Based on the combination of soil and land use
10	Distance from the nearest stream (*Distance*)	m	An important factor in flow transfer to the outlet

of FA in this particular case study. The value of KMO also points to a significant correlation among components.

The characteristics of principal components are presented in Table 3 including eigenvalues, variance proportion, and cumulative variance proportion. Each component whose variance is at least equal to 1 (representing at least 10% of the total variance) is selected as the principal component (Wolock *et al.* 2004). This threshold of the variance is a qualitative criterion and is flexible depending on the subject or user experience. According to Table 3, it is clear that the first three components account for 98.49% of total variance of input variables.

Each variable with the highest loadings for each component was selected as the most effective variable/feature. According to the rotated component matrix (Table 4), *TPI*, aspect, curvature-plan, curvature-profile, and slope are the most effective on the first component that accounts for more than 82.52% of input variance. One of these five variables should be selected as the representative of the first component. Although all five variables are DEM derivatives, *TPI* indicates the tendency of each cell to saturate. Selection of *TPI* is reasonable in terms of runoff generation process. Furthermore, *Distance* and *CN* have the most effect on the second and third components that account for 11.08% and 4.87% of input variance, respectively. In comparison, Dehotin & Braud (2008) considered DEM, *TPI*, and *CN* layers to determine the HHR without preprocessing.

Thus, *TPI*, distance, and *CN* were identified as the most effective surface features of the watershed along with the distributed rainfall layer. The effect of using the third layer (*CN*) on clustering results will be further examined.

SOMFCM clustering

Determining cluster centers

Before performing fuzzy clustering (FCM algorithm), it is necessary to initialize cluster centers. For this purpose, a $6{,}225 \times 4$ matrix, involving three surface features/variables and a rainfall variable, were subjected to SOFM and the results were evaluated based on the silhouette width (Rousseeuw 1987).

A hexagonal lattice was recommended by Kohonen (2001) since it does not favor horizontal or vertical directions. In this study, a matrix of nodes with a rectangular or hexagonal lattice neighborhood structure was used. Therefore, the best configuration of SOFM was evaluated through several runs. The best architecture including the number of clusters (c) and lattice neighborhood structure based on the silhouette width are presented in Table 5.

The best network arrangement for each number of clusters (as presented in Table 5) was entered into the fuzzy clustering. It should be noted that results of using Ward algorithm, though slightly weaker, did not differ significantly with those of the SOFM.

Fuzzy clustering validation indices to determine optimum *c* and *m*

Based on the best type of lattice neighborhood structure (Table 5) corresponding to the different number of clusters ($4 \leq c \leq 11$), the sensitivity to variation of the fuzzification parameter ($1.1 \leq m \leq 3$) with an increment of 0.1 was examined. Pal & Bezdek (1995) reported that the FCM provides better performance for m in 1.5–2.5 range. However, in this study, the SOMFCM hybrid fuzzy clustering was performed in 160 possible runs (i.e., 20 different states for $1.1 \leq m \leq 3$ and eight different states for $4 \leq c \leq 11$) for each proposed scenario. To determine the optimum c, the results of the five selected validation indices introduced in available studies (e.g., Rao & Srinivas 2006, 2008; Odgers *et al.* 2011; Farsadnia *et al.* 2014; Prasad & Arora 2014) are

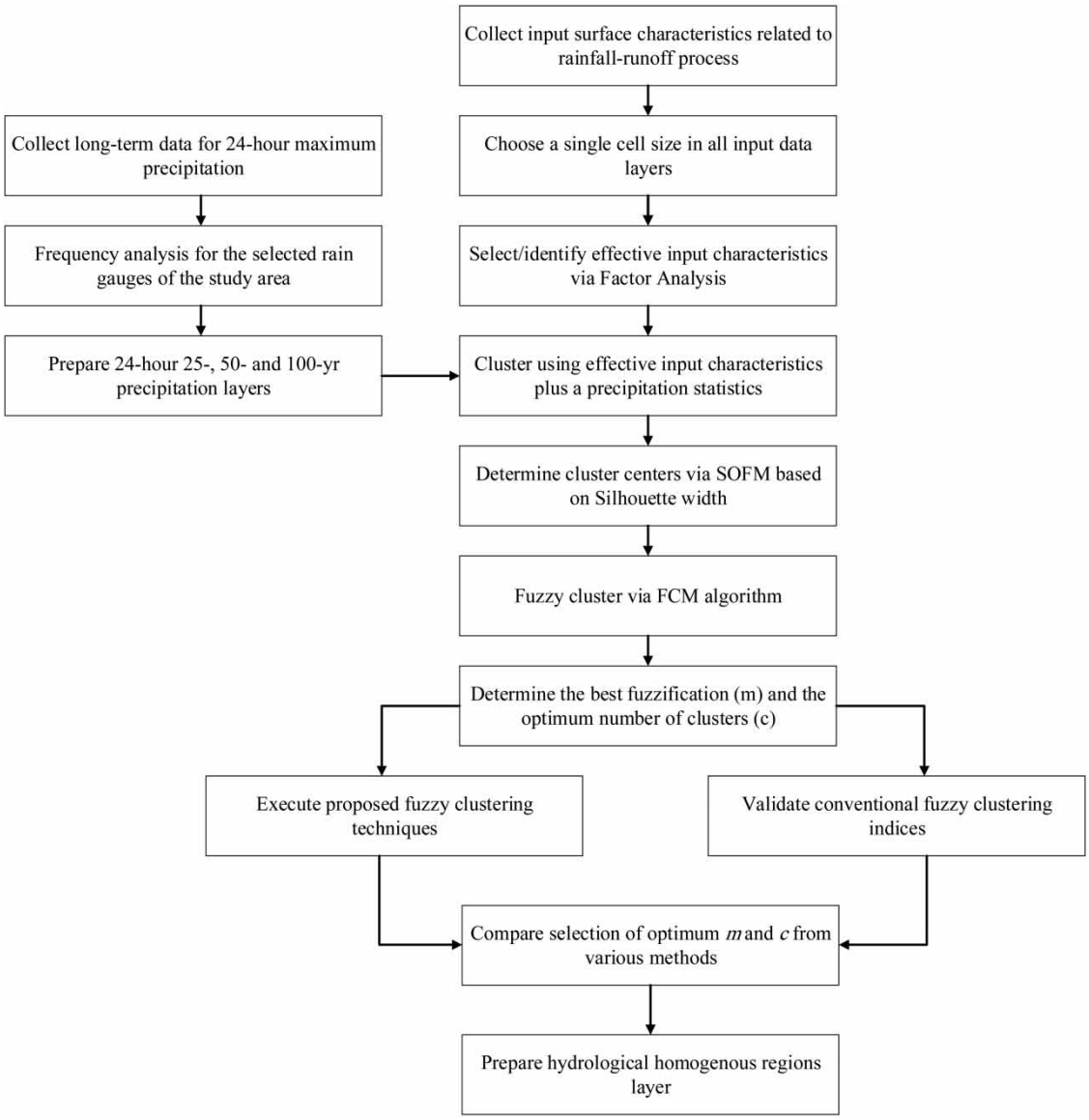

Figure 2 | Flow chart of the study procedure.

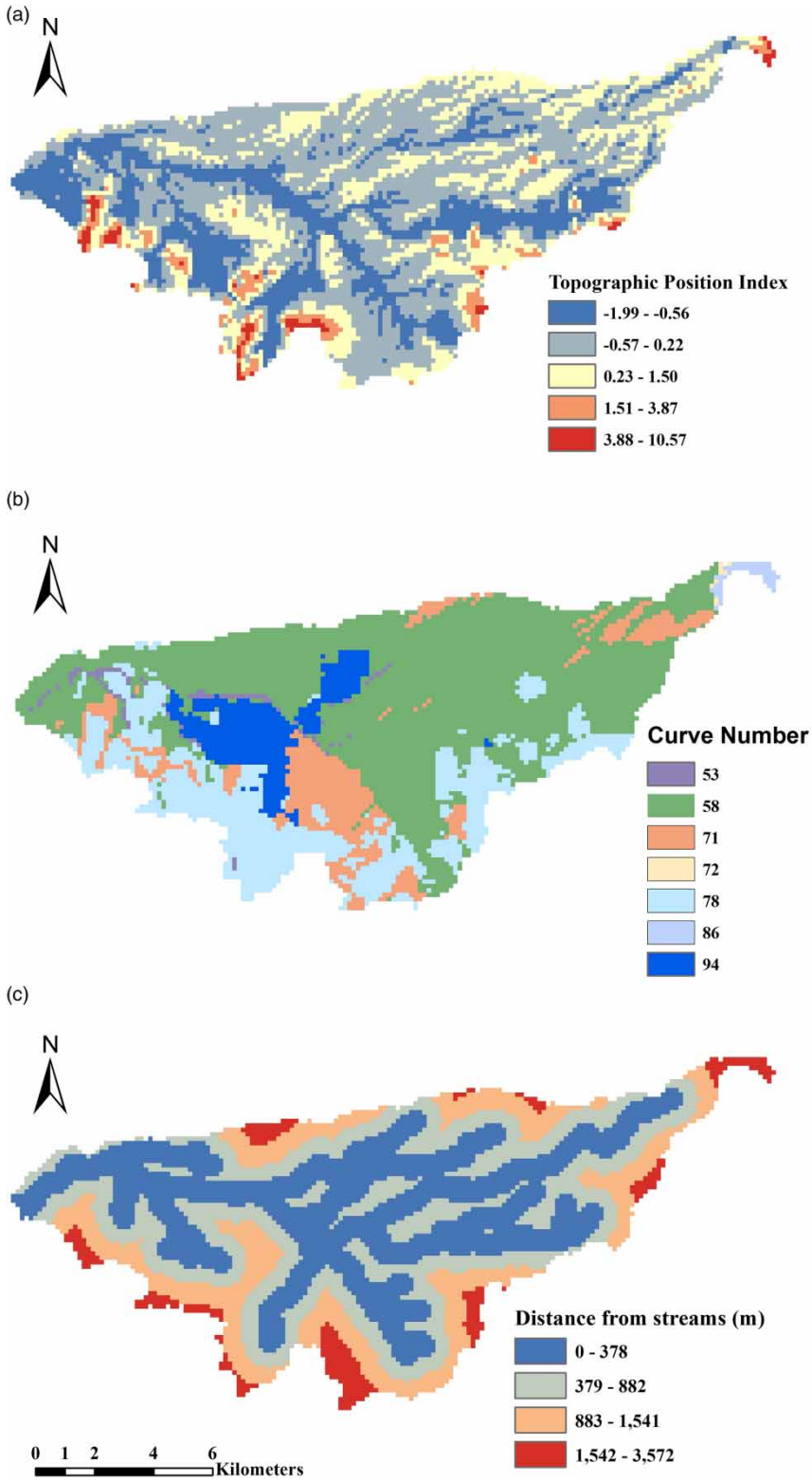

Figure 3 | Some selected input surface feature layers used in factor analysis (cell size = 150 m): (a) topographic position index (*TPI*), (b) curve number (*CN*), and (c) distance from the nearest stream (*Distance*).

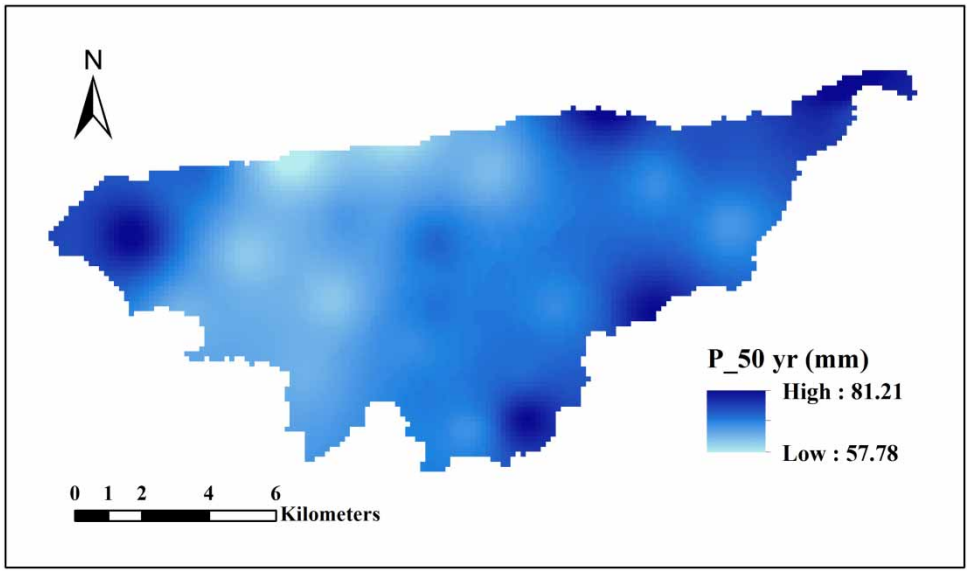

Figure 4 | 50-year rainfall layer (cell size = 150 m).

presented in Table 6. It should be noted that for each clustering scenario, the upper bound of m was reduced due to divergent non-meaningful results. In Table 6, the case of $c = 6$ and $m = 1.6$ is marked in bold type in order to facilitate comparison.

In Table 6, *Euc_mean* is the mean Euclidean distance between the actual feature values of the cells and the ones corresponding to cluster center, R^2 represents the ratio of the variance of the data corresponding to the cluster centers to the variance of the actual data, and *OF* is the objective function. Since for each clustering scenario, a threshold fuzzification parameter (m) corresponding to $N\_Umax = 0$ is defined and the number of cells with a membership degree greater than the specified value in the present study (i.e., $1-1/c$) will be equal to zero ($N\_Umax = 0$), the results of the validation indices are presented to the threshold m. It is also worth noting that, following Chiu (2005), the cells

Table 3 | Descriptive statistics for different components

	Initial eigenvalues		
Component (VAR)	**Total**	**% of variance**	**Cumulative variance %**
1	8.253	82.529	82.529
2	1.108	11.082	93.611
3	0.488	4.876	98.487
4	0.150	1.505	99.991
5	0.001	0.009	100
6	1.392×10^{-5}	0.000	100
7	6.275×10^{-6}	6.275×10^{-5}	100
8	9.385×10^{-7}	9.385×10^{-6}	100
9	8.995×10^{-8}	8.995×10^{-7}	100
10	1.128×10^{-8}	1.128×10^{-7}	100

Note: The first three components were selected and subjected to the varimax normalized rotation (Overall & Klett 1972). This method of rotation has been widely accepted as the most appropriate type of orthogonal rotation (Puvaneswaran 1990; White *et al.* 1991). The most effective variables in component formation are shown in bold in Table 4.

Table 4 | Factor loadings for ten selected variables resulted from rotated component matrix (extraction method: principal component analysis, rotation method: varimax with Kaiser normalization)

	Component		
Variable	**1**	**2**	**3**
VAR1 (NDVI)	0.573	0.801	0.027
VAR2 (Ksat)	0.573	0.801	0.026
VAR3 (TPI)	**0.901**	0.432	−0.026
VAR4 (Aspect)	**0.901**	0.432	−0.026
VAR5 (Curvature- plan)	**0.901**	0.432	−0.026
VAR6 (Curvature-profile)	**0.901**	0.432	−0.026
VAR7 (Slope)	**0.901**	0.432	−0.026
VAR8 (CN)	−0.053	0.084	**0.994**
VAR9 (Distance)	0.433	**0.834**	0.167
VAR10 (DEM)	**0.901**	0.431	−0.028

Table 5 | Best SOFM architecture

Number of clusters (c)	Width	Length	Type of lattice neighborhood structure	Silhouette width
4	4	1	Rectangular/ Hexagonal	0.3737
5	5	1	Rectangular	0.4292
6	6	1	Rectangular	0.4376
7	7	1	Rectangular/ Hexagonal	0.4592
8	4	2	Rectangular	0.4629
9	9	1	Rectangular	0.455
10	5	2	Rectangular/ Hexagonal	0.4538
11	11	1	Hexagonal	0.4538

with a degree of membership less than 1/c are not assigned to any cluster.

The optimum c corresponds to the minimum values of V_{EXB}, V_K, V_{FS}, and V_{PE} as well as the maximum V_{PC}. The values of the selected clustering validation indices in this study (as evident by Table 6) indicate that three-component clustering scenario outperforms the four-component clustering scenario. Accordingly, the fourth component or third surface feature (i.e., *Distance* in this study) did not improve the results. Figure 5 shows the variations of the selected validation indices for different values of m and c corresponding to the three-component clustering (*TPI*, *CN*, and *P50*).

As shown in Figure 5(a)–5(c), it is not possible to determine optimum c and m based on V_{PC}, V_{PE}, and V_{FS}. The V_{FS} follows a slow decreasing trend as the number of clusters increases over 8 while the value of V_{PC} has a monotonic decreasing trend. However, the V_{PE} has a monotonic increasing trend as the number of clusters increases. Overall, the indices (V_{PC} and V_{PE}) suggest $c = 4$ as the optimum c in this study. This is because neither V_{PC} and V_{PE} have a direct connection to any property of the input data (Zhang *et al.* 2008) and cannot solely serve as a reasonable basis for choosing optimum c. Consequently, these indices seem to be more appropriate when well-separated clusters with crisp boundaries are formed.

On the other hand, as shown in Figure 5(d) and 5(e), the trends of V_k and V_{EXB} are similar so that the minimum values obtained from these indices may be the basis for choosing $c = 7$ clusters while m falls in the 1.5–1.7 range. However, V_{FS}, V_{EXB}, and V_k variations show monotonic decreasing or increasing the tendency for majority m and c values, even though there are a few exceptions. Hence, no decisive judgment is expected on the optimum c on the basis of these indices. Our results are similar to those of Rao & Srinivas's (2006) study in that the selected indices may lead to the poor discretization of clusters. Therefore, it is necessary to determine optimum c and m

Table 6 | Range of selected validation indices in various clustering scenarios

						Technique I		Technique II		Technique III	
Clustering scenario	**Index**	V_{EXB}	V_k	V_{FS}	**OF**	V_{PC}	V_{PE}	***CI_mean***	***Euc_mean***	N_U_{max}	R^2
1. Four-component (*TPI-Distance-CN-P25*)	min ($1.1 < m < 2.2$)	0.495	3,088	−21,395	1,295	0.205	0.083	0.058	1.762	5	0.264
	max ($1.1 < m < 2.2$)	1.785	11,110	−7,544	14,515	0.951	1.723	0.834	1.936	5,850	0.748
	for c = 6 and m = 1.6	**0.628**	**3,914**	**−10,378**	**6,586**	**0.574**	**0.874**	**0.460**	**1.886**	**2,061**	**0.598**
2. Four-component (*TPI-Distance-CN-P50*)	min ($1.1 < m < 2.6$)	0.498	3,099	−19,217	1,296	0.188	0.075	0.056	1.741	1	0.160
	max ($1.1 < m < 2.6$)	2.339	14,544	−1,096	12,191	0.957	2.044	0.907	1.936	5,866	0.751
	for c = 6 and m = 1.6	**0.618**	**3,851**	**−10,109**	**6,619**	**0.563**	**0.892**	**0.477**	**1.882**	**1,976**	**0.593**
3. Four-component (*TPI-Distance-CN- P100*)	min ($1.1 < m < 2.3$)	0.503	3,127	−21,166	1,036	0.284	0.099	0.062	1.762	2	0.264
	max ($1.1 < m < 2.3$)	2.569	16,170	−7,926	12,184	0.943	1.757	0.871	1.944	5,861	0.760
	for c = 6 and m = 1.6	**0.691**	**4,306**	**−10,395**	**6,649**	**0.563**	**0.888**	**0.487**	**1.902**	**1,915**	**0.596**
4. Three-component (*TPI-CN-P50*)	min ($1.1 < m < 2.7$)	0.171	1,166	−17,259	280	0.257	0.055	0.027	1.516	8	0.408
	max ($1.1 < m < 2.7$)	0.659	4,154	−2,752	7,411	0.968	1.895	0.812	1.670	6,456	0.839
	for c = 6 and m = 1.6	**0.386**	**2,416**	**−10,509**	**3,896**	**0.690**	**0.642**	**0.326**	**1.636**	**3,261**	**0.739**
Previous studies	min	0.13	−602.51	81.87	160	0.99	0.05	0	–	–	–
	max	14.49	852.36	3,549.6	1,280	0.21	2.11	1	–	–	–

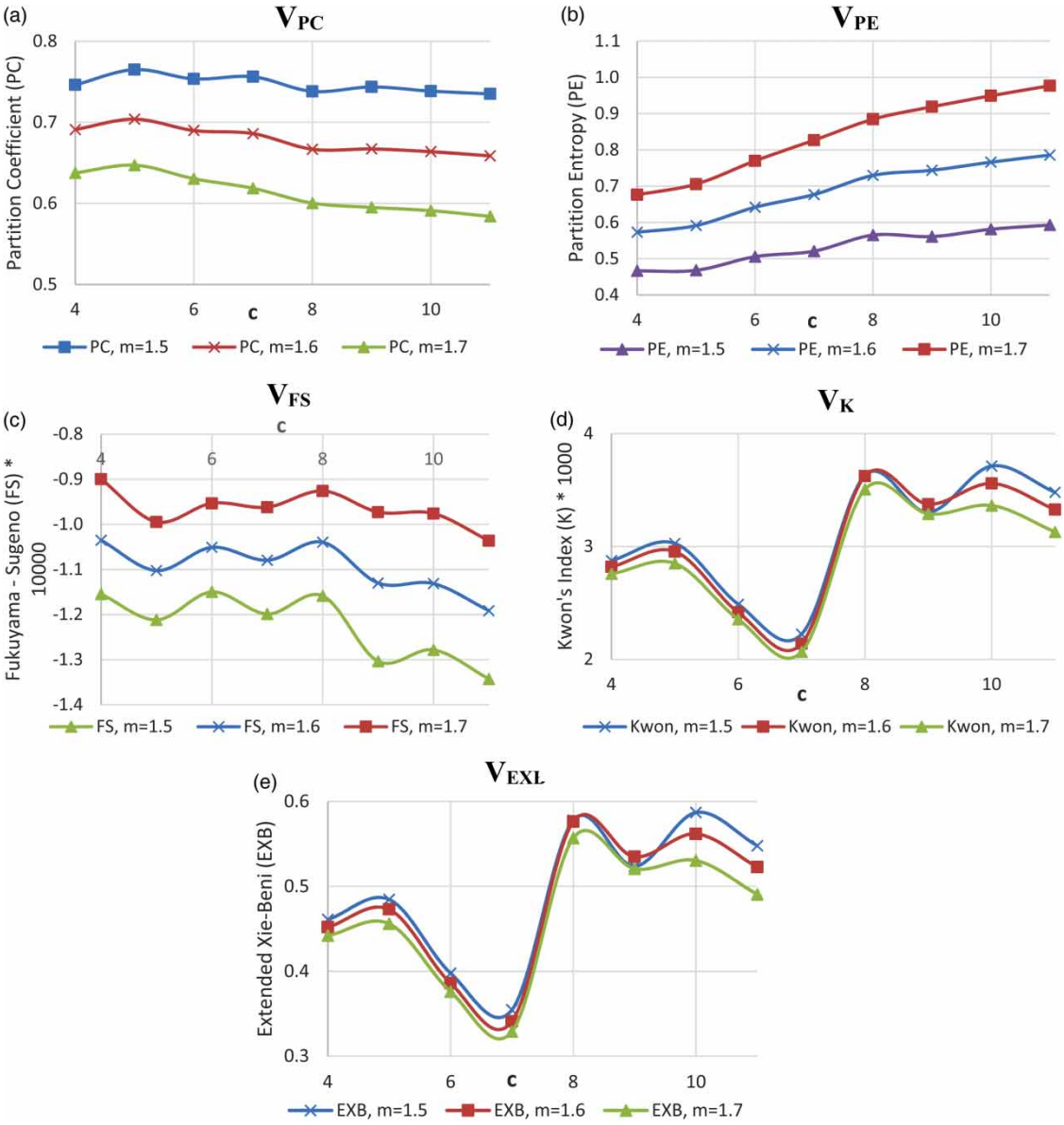

Figure 5 | Variations of selected fuzzy validation indices for clustering scenario 4 (TPI-CN-P50): (a) V_{PC}, (b) V_{PE}, (c) V_{FS}, (d) V_K, and (e) V_{EXB}.

through other means. Based on Table 6, one may infer the following:

- The range of R^2 varies from 0.16 to 0.76 for all four-component clustering scenarios and from 0.41 to 0.84 for the three-component clustering scenario. This indicates that clustering with three input components is better than the four input components.
- Comparing the mean *CI* between the three- and four-component clustering scenarios for rainfall shows that the three-component scenario involving 50-year rainfall is more reliable (mean *CI* values in the three-component scenario are less than the ones corresponding to all four-component scenarios for different return periods).
- For all clustering scenarios, the number of cells (*N_Umax*) belonging to the clusters with a certain degree of membership (as $\alpha_{high} \geq 1-1/c$) has a decreasing trend with increasing m and c.
- *N_Umax* for clustering scenario 4 was greater than the ones in other clustering scenarios for all values of m

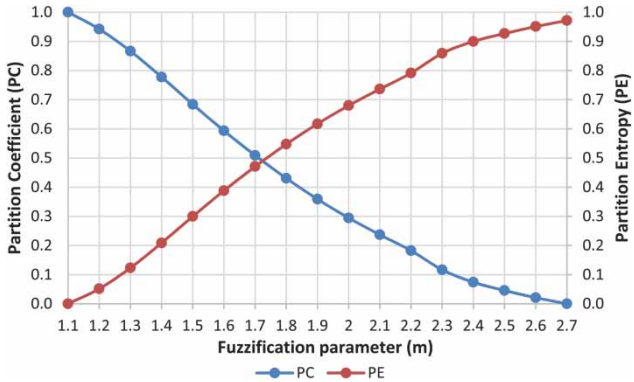

Figure 6 | Interaction between PC and PE (in technique I) for $c = 6$ (clustering scenario 4: TPI-CN-P50).

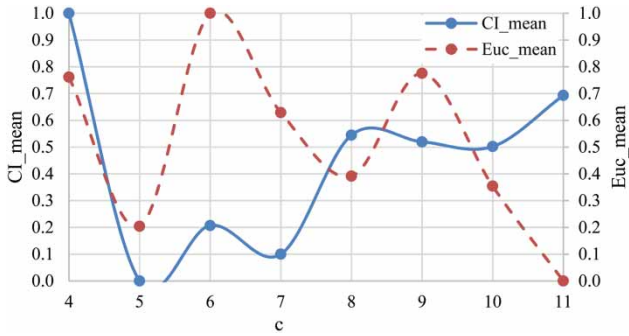

Figure 7 | Interaction between CI_mean and Euc_mean (technique II) for $m = 1.7$ (clustering scenario 4: TPI-CN-P50).

and c, which suggests that the clusters (regions) may be obtained without the distance layer. Therefore, distance layer has no significant effect on the results and is removed in further analysis.

Proposed validation techniques for determining optimum *c* and *m*

Due to the inadequacy of conventional validation indices (Table 6 and Figure 5), three validation techniques are proposed for determining the optimum c and m. In all three techniques, the standardization of the inputs in [0,1] interval was first performed for both vertical axes.

Technique I of determining optimum m: The results of the application of technique I corresponding to $c = 6$ are shown in Figure 6.

Similar results, as shown in Table 7, were obtained for the different clustering scenarios and different number of clusters ($4 \le c \le 11$). In most cases of c, the point of V_{PC} and V_{PE} intersection occurred at $m = 1.6$ (except Figure 6). Therefore, technique I suggests the optimum $m = 1.6$ for most scenarios.

Technique II of determining optimum c: Figure 7, as an example, shows the interaction between *CI_mean* and *Euc_mean* for optimum $m = 1.7$.

As shown in Figure 7, there are three intersections near c equal to 5, 8, and 10. The criterion for determination of optimum c is the highest frequency of intersections. Therefore, by drawing the interactions similar to Figure 7 for different m and c and for different clustering scenarios, the optimum c may be determined. The results are presented in Table 7.

Technique III of determining optimum c: As an example, Figure 8 shows the interaction for $m = 1.7$ (as optimum m).

As shown in Figure 8, in this case, the intersection occurred at only one point, close to $c = 6$. In technique III, there is one intersection point for each interaction while the criterion for optimum is the highest frequency of intersections for different interactions. Therefore, by drawing the interactions similar to Figure 8 for different m and c, and for different clustering scenarios, optimum c may be obtained as presented in Table 7.

Both optimum c and m were selected based on the highest frequency for intersection points. There were two

Table 7 | Final results of determination optimum c and m for different clustering scenarios

Clustering scenario (input components)	Range of *m*	Optimum fuzzification parameter (*m*) Technique I	Optimum number of cluster (*c*) Technique II	Technique III
Scenario 1 (TPI-CN-Distance-P50)	$1.1 \le m \le 2.2$	1.6	7	6
Scenario 2 (TPI-CN-Distance-P25)	$1.1 \le m \le 2.6$	1.6	8	6
Scenario 3 (TPI-CN-Distance-P100)	$1.1 \le m \le 2.3$	1.6	4	6
Scenario 4 (TPI-CN-P50)	$1.1 \le m \le 2.7$	1.7	6	6

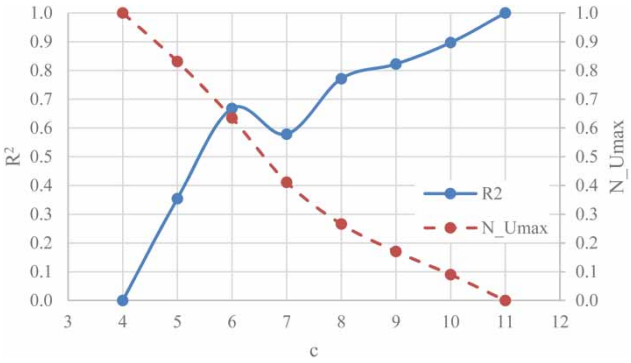

Figure 8 | Interaction between R^2 and N_Umax (technique III) for $m = 1.7$ (clustering scenario 4: TPI-CN-P50).

reasons for limiting upper bound of m for each clustering scenario: (1) a threshold for m corresponding to $N\_Umax = 0$ and (2) the mean threshold for CI which is equal to 0.6. The results were subjected to these thresholds.

Based on Table 7, the highest priority corresponds to $c = 6$. In other words, the WGEW watershed may be divided into six HHR. Therefore, out of 160 different selections in SOMFCM, the optimum values of $m = 1.6$ and $c = 6$ predominated in most clustering scenarios. Now, the best clustering scenario should be selected.

Other interpretations derived from Tables 6 and 7 are as follows:

- Techniques I (optimum m) and III (optimum c) yield similar results in different clustering scenarios. The results of technique II are somewhat different, because of variations in mean CI and mean Euclidean distance for different values of m.
- Technique III in all clustering scenarios presents similar results, which indicates the accuracy of the selected clustering features.
- Rainfall return periods only affect technique II, especially *CI_mean* and *Euc_mean* parameters. However, the results of techniques I and III are similar in all return periods.
- It may be concluded that in the present study, the distance from the nearest channel (distance layer) has no significant effect on optimum m and c. In other words, clustering can be performed with two more effective land layers, namely, TPI and CN. As such, clustering requires less data, which is desirable in the delineation of HHR.

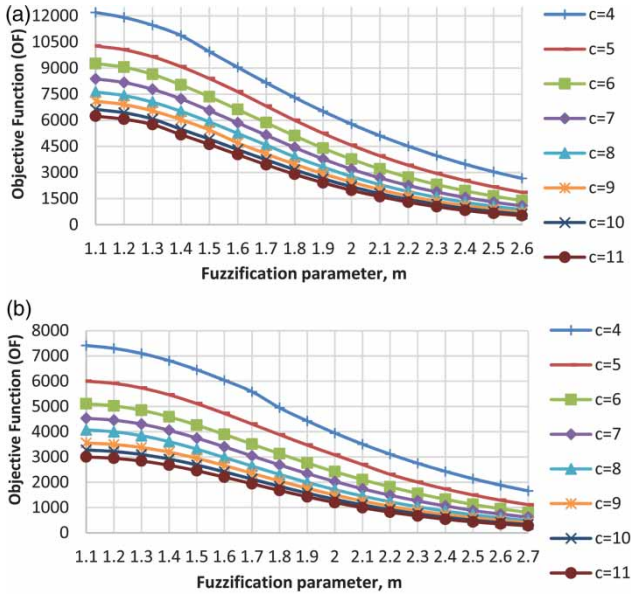

Figure 9 | Variation of objective function (*OF*) value for different m and c: (a) fuzzy clustering scenario 1 and (b) fuzzy clustering scenario 4.

To better compare two superior clustering scenarios (i.e., scenarios 1 and 4) as far as the distance layer is concerned, the changes in optimal OF value are presented in Figure 9.

Generally speaking, the OF value reduces when c increases for a specified value of m. According to Figure 9, after the removal of the distance layer as the third surface feature, the OF value is significantly reduced. Therefore, in order to achieve minimum OF in FCM algorithm, two effective surface features are adequate for clustering.

Determination of hydrologic homogenous regions

Figure 9 and Tables 6 and 7 indicate that clustering scenario 4 involving TPI, CN, and P50 is the best. Clustering maps corresponding to the best three-component and four-component scenarios are shown in Figures 10 and 11. Although $m = 1.7$ was optimum in the three-component scenario based on technique I, other indices, such as mean CI and mean Umax, were better produced for $m = 1.6$ (second rank in technique I). Statistical comparison between cluster cells and cluster centers, as well as MARD, are presented in Tables 8 and 9 for four-component and three-component clustering scenarios, respectively.

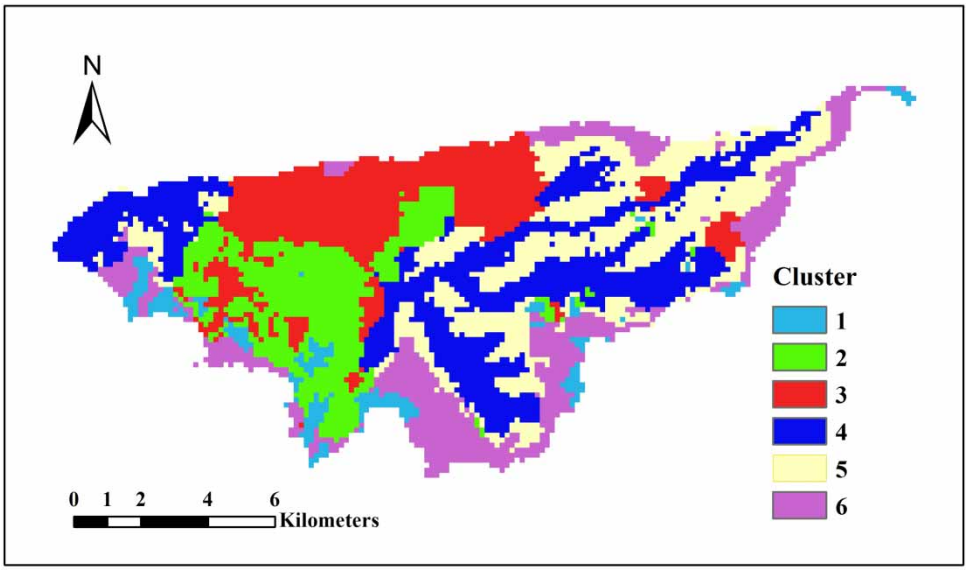

Figure 10 | Fuzzy clustering map for four-component clustering scenario (TPI-CN-Distance-P50) for $m = 1.6$ and $c = 6$.

In Tables 8 and 9, *MARD* is the mean absolute relative distance which is determined as follows:

$$MARD = \frac{1}{f}\sum_{i=1}^{f}\left|\frac{(\bar{x}_{if} - \bar{\bar{c}}_{if})}{\bar{x}_{if}}\right| \quad (14)$$

where f is the number of selected components (in this study, f = 3 for three-component and 4 for four-component clustering), $\bar{x}_{if}$ and $\bar{\bar{c}}_{if}$ are the mean and cluster center values of the f-feature within cluster I, respectively. Therefore, with respect to the MARD values, SOMFCM clustering has produced valid maps in both clustering scenarios.

As presented in Tables 8 and 9, comparison of features cell values with those of cluster centers indicates that there is a significant similarity between cells and cluster

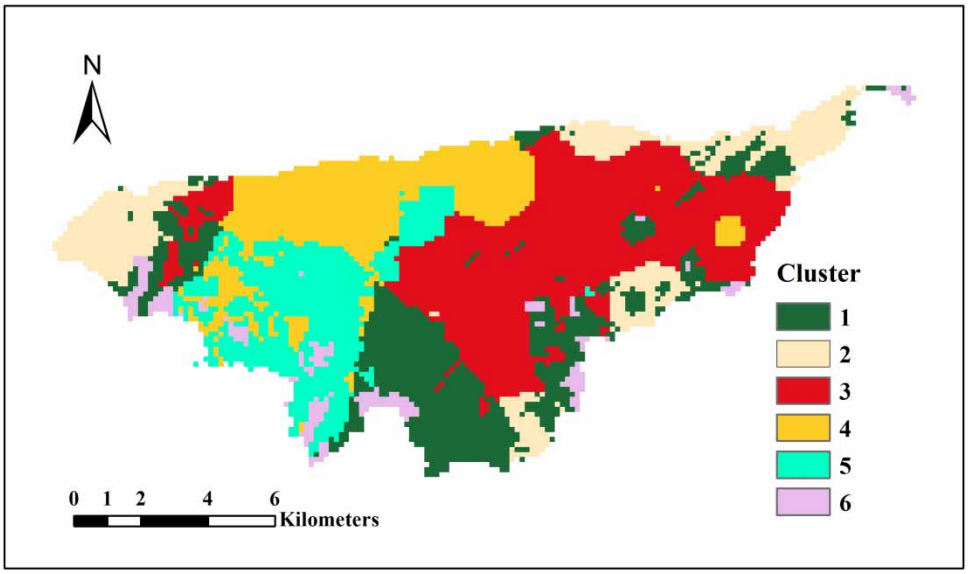

Figure 11 | Fuzzy clustering map for three-component clustering scenario (TPI-CN-P50) for $m = 1.6$ and $c = 6$.

Table 8 | Four-component fuzzy clustering map statistics ($m = 1.6$ and $c = 6$)

					Cluster cells				Cluster centers				
Cluster	Number of cells	Area (%)	Mean CI	Mean Umax	Mean TPI	Mean CN	Mean *Distance* (m)	Mean P50 (mm)	Mean TPI	Mean CN	Mean *Distance* (m)	Mean P50 (mm)	MARD (%)
1	276	4.44	0.400	0.706	3.53	77.99	1,223.39	68.13	3.47	77.65	1,187.92	67.95	1.30
2	891	14.83	0.410	0.708	−0.24	87.37	482.49	65.45	−0.22	89.04	468.42	65.57	3.50
3	849	18.99	0.569	0.609	0.26	72.56	1,358.51	68.58	0.25	72.74	1,290.40	69.14	2.81
4	1,227	24.72	0.422	0.729	−0.18	59.11	473.41	65.15	−0.17	59.63	444.54	65.54	3.20
5	2,090	20.89	0.449	0.715	−0.38	60.38	249.90	70.48	−0.39	59.59	261.72	70.32	2.21
6	921	16.14	0.614	0.595	0.036	63.14	797.02	74.08	0.03	62.77	760.16	73.50	3.88
Total	6,210	100	0.477	0.684	0.504	70.09	764.12	68.64	0.496	70.24	735.53	68.67	2.81

centers in each cluster. MARD values imply that four-component clustering error is less than that of the three–component clustering. This is because in four-component clustering, the addition of one input component to the clustering procedure improves the accuracy of assigning cells to the clusters. On the other hand, adding one more component increases the computation time and a relative error reduction of 10.22% (3.13–2.81/3.13). This error is negligible, so the third surface feature (distance layer) has no significant effect on the results.

As shown in Figure 11, in addition to its remarkable similarity with the CN map shown in Figure 2, the discretized regions (clusters) also identify the river network, plains, and man-made areas.

Furthermore, Tables 8 and 9 indicate that the area of the smallest cluster is 4.44% (6.66 km^2) and 4% (6 km^2) corresponding to the four-component and three-component clustering scenarios, respectively. Such values, as reported by Boulaine (1980), should be less than the smallest distinguishable area on a map (expressed in this study as the threshold area) that is equal to $0.5\ cm \times 0.5\ cm = 0.25\ cm^2$ on the map. With scale 1:24,000 used in this study, the threshold area is $120\ m \times 120\ m = 14{,}400\ (m^2)$, almost 1.44% of the total basin area (2.16 km^2). Therefore, the aforementioned smallest areas for both clustering scenarios are larger than the threshold area.

The *CI_mean* values were obtained as 0.477 and 0.326, indicating that the uncertainty of clusters falls in low and very low categories for the four-component and three-component clustering scenarios, respectively. The average degree of membership is equal to 0.684 for the four-component clustering and 0.788 for the three-component clustering scenarios, which are slightly less than $\alpha_{High} = (1–1/6) = 0.833$. These results indicate that the three-component clustering map is more reliable and accurate than the four-component map.

Table 9 | Three-component fuzzy clustering map statistics ($m = 1.6$ and $c = 6$)

					Cells			Cluster centers			
Cluster	Number of cells	Area (%)	Mean CI	Mean Umax	Mean TPI	Mean CN	Mean P50 (mm)	Mean TPI	Mean CN	Mean P50 (mm)	MARD (%)
1	1,243	19.97	0.398	0.731	−0.063	74.81	70.74	−0.07	74.47	70.50	2.28
2	821	13.19	0.399	0.755	−0.281	59.15	74.86	−0.29	59.53	74.54	1.20
3	1,930	31.00	0.274	0.829	−0.165	57.93	69.66	−0.17	58.39	69.68	0.39
4	1,064	17.09	0.243	0.848	−0.210	59.01	64.48	−0.18	59.13	64.57	5.12
5	918	14.75	0.366	0.749	−0.168	87.10	64.76	−0.14	89.56	64.98	8.49
6	249	4.00	0.337	0.746	3.472	77.97	68.34	3.61	77.87	68.34	1.31
Total	6,225	100	0.326	0.788	0.431	69.33	68.81	0.46	69.83	68.77	3.13

Table 10 | Comparison between the best SOMFCM clustering scenarios based on conventional fuzzy validation indices and proposed techniques

The best clustering scenario	Clustering validation type	Optimum *m*	Optimum *c*	V_{EXB}	V_K	V_{FS}	MARD (%)	CI_{mean}	$(Umax)_{mean}$	*N_Umax*	R^2
Four-component (*TPI-CN-Distance-P50*)	Conventional index	1.7	8	0.516	3,225	−8,738	3.25	0.535	0.615	938	0.632
Four-component (*TPI-CN-Distance-P50*)	**Proposed technique**	**1.6**	6	0.618	3,851	−10,109	2.81	0.477	0.684	1,976	0.593
Three-component (*TPI-CN-P50*)	Conventional index	1.7	7	0.329	2,266	−9,620	15.31	0.38	0.739	2,266	0.716
Three-component (*TPI-CN-P50*)	**Proposed technique**	**1.6**	6	0.386	2,416	−10,509	3.13	0.326	0.788	3,261	0.739
		1.7	6	0.376	2,355	−9,533	3.51	0.385	0.731	2,686	0.737
Three-component (*TPI-Distance-P50*)	**Proposed technique**	**1.6**	6	0.647	4,184	−9,975	3.53	0.405	0.738	2,659	0.657

The results of the best clustering scenarios with different inputs derived from the conventional fuzzy clustering validation indices along with the proposed techniques are presented in Table 10.

According to Table 10, the three-component clustering scenario (TPI-CN-P50), shown in bold, relies on the validation techniques proposed in the study and presents the most reliable clustering map (Figure 11). Some physical justifications are as follows:

- By comparing the final cluster map (Figure 11) with the 50-year rainfall map (Figure 4), it is evident that clusters 1 and 2 (together covering 43% of the watershed area), and clusters 4 and 5 (together covering 32% of the watershed area) correspond to areas which receive maximum and minimum 50-year rainfall, respectively. It is worth noting that the dominant rainfall occurs in the eastern half of the WGEW (within clusters 1, 2, and 3).
- By comparing the final cluster map (Figure 11) with the CN map (Figure 3(b)) it is indicated that cluster 5 is approximately representative of maximum CN areas. Clusters 2, 3, and 4 (61% of the watershed area) cover areas with minimum CN.
- No clear pattern may be observed between the final cluster map (Figure 11) with the *TPI* and *Distance* maps (Figure 3(a) and 3(c)).

SUMMARY AND CONCLUSIONS

The objective of this study was to provide an efficient methodology for identification/delineation of hydrologic homogeneous regions (HHR) in the context of the rainfall-runoff framework. For this purpose, a hybrid fuzzy clustering method, namely SOMFCM, was adopted. Specifying initial cluster centers in the FCM algorithm affects the final clusters. In this paper, a hybrid approach was proposed to improve the performance of FCM algorithm for partitioning of watersheds. SOMFCM is a combination of SOFM (for determining the initial centers and the location of fuzzy clusters) and the robust and the well-known FCM algorithm (for watershed discretization). At first, a number of attributes/components associated with rainfall-runoff processes were identified and mapped. The components consisted of ten

physiographic/surface features and one of three rainfall values (with return periods of 25, 50 and 100 years) representing a meteorological factor. All maps were prepared in 150-meter cell size. All surface features were entered into the factor analysis (FA) in order to reduce the dimensions of the input layers. The most effective features were identified as TPI, CN, and *Distance* that encompass physiographical and soil/vegetation attributes within the clustering. Therefore, these three layers along with one of the three rainfall layers were forwarded into the SOMFCM clustering procedure.

SOMFCM cluster maps were produced for $1.1 \leq m \leq 3$ and $4 \leq c \leq 11$ in two clustering scenarios involving four-component (with distance layer) and three-component cases (without distance layer). Then, to validate the SOMFCM clustering results, in addition to conventional and well-known fuzzy validation indices, a number of techniques for determining the optimum number of clusters (c) and fuzzification parameter (m) were introduced. The main conclusions of this study are as follows:

- Although using the maximum degree of membership is a common way to determine the winning cluster, an acceptable level of α-*cut* is required. In this study, the criterion of (α-$cut = 1 - 1/c = 0.833$, for $c = 6$) was applied. Although α-$cut = 0.833$ value was not met for all cells, best results were obtained for the three-component clustering scenario (TPI-CN-P50). The number of clusters was determined in such a way that the mean degree of membership of all cells was 0.788 (slightly less than 0.833), while 3,261 cells (52% of all cells) had a degree of membership more than 0.833. The choice of this criterion for assigning the cells into the clusters in the context of fuzzy clustering methods appears logical. The *CI_mean* is another validation index that turned out to be 0.326, which is the minimum value among the results of other clustering scenarios as well as within the range of very low uncertainty ($CI \in (0.2–0.4)$).
- Application of the selected conventional validation indices (V_{PC}, V_{PE}, V_{FS}, V_{EXB}, and V_K) and the proposed validation techniques showed that among 160 different clustering runs for each scenario, optimum clustering involves six clusters ($c = 6$) with fuzzification parameter (m) of 1.6, derived from the *TPI*, *CN*, and *P50* features.
- It became clear that use of conventional validation indices, despite having a strong mathematical and rational theory, will not necessarily lead to an optimum solution with acceptable performance in the delineation of HHRs. On the other hand, the techniques introduced in the study may be more effective in determining the optimum number of hydrologically homogeneous clusters.
- The performance of FCM algorithm was improved when coupled with SOFM. In general, the quality of clustering is improved via coupled SOFM-FCM. Therefore, use of the SOMFCM with a fewer number of input layers and properly prioritized via factor analysis is recommended in clustering the HHRs.

Comparison of the HHR map resulted from clustering for WGEW with a distributed map showing the basin runoff to provide a proper judgment on the effectiveness of the clustering practice is suggested for future research. Also, the application and evaluation of the proposed methodology on other watersheds using the more effective layers introduced in this study can also be suggested. In addition, the assessment of other effective watershed features and the optimal combination of various features in preparation of a clustering map along with sensitivity analysis of SOMFCM with respect to DEM cell size may be conducted in future research.

REFERENCES

Ahani, A. & Nadoushani, S. S. M. 2016 Assessment of some combinations of hard and fuzzy clustering techniques for regionalization of catchments in Sefidroud basin. *Journal of Hydroinformatics* **18** (6), 1033–1054.

Bartzokas, A., Lolis, C. J. & Metaxas, D. A. 2003 A study of the intraannual variation and the spatial distribution of precipitation amount and duration over Greece on a 10-day basis. *International Journal of Climatology* **23**, 207–222.

Basu, B. & Srinivas, V. V. 2015 Analytical approach to quantile estimation in regional frequency analysis based on fuzzy framework. *Journal of Hydrology* **524**, 30–43.

Bezdek, J. C. 1974a Numerical taxonomy with fuzzy sets. *Journal of Mathematical Biology* **1**, 57–71.

Bezdek, J. C. 1974b Cluster validity with fuzzy sets. *Journal of Cybernetics* **3**, 58–74.

Bezdek, J. C. 1981 *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York.

Bloschl, G. & Sivapalan, M. 1995 Scale issues in hydrological modelling: a review. *Hydrological Processes* **9**, 251–290.

Boulaine, J. 1980 *Applied Pedology*. Agricultural Sciences Collection, Masson, Paris.

Bruin, S. & Stein, A. 1998 Soil-landscape modeling using fuzzy c-means clustering of attributes data derived from a Digital Elevation Model (DEM). *Geoderma* **83**, 17–33.

Burrough, P. A., van Gaans, P. F. M. & Hootsmans, R. 1997 Continuous classification in soil survey: spatial correlation, confusion and boundaries. *Geoderma* **77**, 115–135.

Chiu, W. Y. 2005 *Wetland Mapping Through Semivariogram Guided Fuzzy Segmentation of Multispectral Imagery*. MSc Thesis, University of Calgary, Alberta, Canada.

Dehotin, J. & Braud, I. 2008 Which spatial discretization for distributed hydrological models? Proposition of a methodology and illustration for medium to large-scale catchments. *Hydrology and Earth System Sciences* **12** (3), 769–796.

Farsadnia, F., Rostami, K., Moghaddam, M., Modarres, R., Bray, M. T., Han, D. & Sadatinejad, J. 2014 Identification of homogeneous regions for regionalization of watersheds by two-level self-organizing feature maps. *Journal of Hydrology* **509**, 387–397.

Ferraro, M. B. & Giordani, P. 2015 A toolbox for fuzzy clustering using the R programming language. *Fuzzy Sets and Systems* **279**, 1–16.

Flügel, W. A. 1995 Delineating hydrological response units by geographical information system analyses for regional hydrological modeling using PRMS/MMS in the drainage basin of the river Brol in Germany. *Hydrological Processes* **9** (3–4), 423–436.

Fukuyama, Y. & Sugeno, M. 1989 A new method of choosing the number of clusters for the fuzzy c-means method. In: *Proceedings of the Fifth Fuzzy Systems Symposium*, Ankara, Turkey, pp. 247–250.

Golian, S., Saghafian, B., Sheshangosht, S. & Ghalkhani, H. 2010 Comparison of classification and clustering methods in spatial rainfall pattern recognition at Northern Iran. *Theoretical and Applied Climatology* **102**, 319–329.

Goodrich, D., Keefer, T., Unkrich, C., Nichols, M., Osborn, H., Stone, J. & Smith, J. 2008 Long-term precipitation database, walnut gulch experimental watershed, Arizona, United States. *Water Resources Research* **44**, W05S04.

Govindaraju, R. S. & Rao, A. R. (eds) 2000 *Artificial Neural Networks in Hydrology*. Kluwer Academic Publishers, The Netherlands, p. 329.

Harman, H. H. 1976 *Modern Factor Analysis*, 3rd edn. The University of Chicago Press, Chicago and London.

Hathaway, R. J. & Bezdek, J. C. 2001 Fuzzy c-means clustering of incomplete data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B* **31**, 735–744.

Heilman, P., Nichols, M. H., Goodrich, D. C., Miller, S. N. & Guertin, D. P. 2008 Geographic information systems database, walnut gulch experimental watershed, Arizona, United States. *Water Resources Research* **44**, W05S11.

Irwin, S. E. 2015 *Assessment of the Regionalization of Precipitation in Two Canadian Climate Regions: A Fuzzy Clustering Approach*. MSc Thesis, University of Western Ontario.

Jenness, J. 2006 Topographic Position Index (tpi_jen.avx) Extension for Arc View 3.x., http://www.jennessent.com.

Keefer, O., Moran, M. S. & Paige, G. B. 2008 Long-term meteorological and soil hydrology database, walnut gulch experimental watershed, Arizona, United States. *Water Resources Research* **44**, W05S07. doi:10.1029/2006WR005702.

Kohonen, T. 1982 Self-organized formation of topologically correct feature maps. *Biological Cybernetics* **43**, 59–69.

Kohonen, T. 2001 *Self-Organizing Maps*. Springer, Berlin, Germany.

Kuentz, A., Arheimer, B., Hundecha, Y. & Wagener, T. 2017 Understanding hydrologic variability across Europe through catchment classification. *Hydrology and Earth System Sciences* **21** (6), 2863–2879.

Kwon, S. H. 1998 Cluster validity index for fuzzy clustering. *Electronics Letters* **34** (22), 176–2177.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. 2015 *Cluster: Cluster Analysis Basics and Extensions*. R package, version 2.0.3.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F. 2015 e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-7. http://CRAN.R-project.org/package = e1071.

Nourani, V. & Parhizkar, M. 2013 Conjunction of SOM-based feature extraction method and hybrid wavelet-ANN approach for rainfall–runoff modeling. *Journal of Hydroinformatics* **15** (3), 829–848.

Nourani, V., Alami, M. T. & Daneshvar, F. 2015 Self-organizing map clustering technique for ANN-based spatiotemporal modeling of groundwater quality parameters. *Journal of Hydroinformatics* **18** (2), 288–309.

Nourani, V., Alami, M. T. & Daneshvar, F. 2016 Hybrid of SOM-clustering method and wavelet-ANFIS approach to model and infill missing groundwater level data. *Journal of Hydrologic Engineering* **21** (9), 05016018.

Odgers, N. P., McBratney, A. B. & Minasny, B. 2011 Bottom-up digital soil mapping. I. Soil layer classes. *Geoderma* **163**, 38–44.

Overall, J. E. & Klett, C. J. 1972 *Applied Multivariate Analysis*. McGraw-Hill, New York.

Pal, N. R. & Bezdek, J. C. 1995 On cluster validity for fuzzy cmeans model. *IEEE Transactions on Fuzzy Systems* **3**, 370–379.

Prasad, M. S. G. & Arora, K. M. 2014 Assessing uncertainty in fuzzy land cover classification by confusion index. *International Journal of Geomatics and Geosciences* **5** (2), 332–344.

Puvaneswaran, M. 1990 Climatic classification for Queensland using multivariate statistical techniques. *International Journal of Climatology* **10**, 591–608.

Rao, A. R. & Srinivas, V. V. 2006 Regionalization of watersheds by fuzzy cluster analysis. *Journal of Hydrology* **318** (1–4), 57–79.

Rao, A. R. & Srinivas, V. V. 2008 *Regionalization of Watersheds – An Approach Based on Cluster Analysis (Water Science and Technology Library)*, Vol. 58. Springer Science + Business Media B. V., The Netherlands, pp. 248.

Reggiani, P., Sivapalan, M. & Hassanizadeh, S. M. 1998 A unifying framework for watershed thermodynamics: balance equations for mass, momentum, energy and entropy, and the second law of thermodynamics. *Advances in Water Resources* **22** (4), 367–398.

Reggiani, P., Sivapalan, M., Hassanizadeh, S. M. & Gray, W. G. 1999 A unifying framework for watershed thermodynamics: constitutive relationships. *Advances in Water Resources* **43** (1), 53–66.

Reggiani, P., Sivapalan, M. & Hassanizadeh, S. M. 2000 Conservation equations governing hillslope responses: exploring the physical basis of water balance. *Water Resources Research* **36** (7), 1845–1863.

Rencher, A. C. 1995 *Methods of Multivariate Analysis*. Wiley, New York.

Rousseeuw, P. J. 1987 Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65.

Saghafian, B. & Khosroshahi, M. 2005 Unit response approach for priority determination of flood source areas. *ASCE Journal of Hydrologic Engineering* **25**, 270–277.

Sharghi, E., Nourani, V., Soleimani, S. & Sadikoglu, F. 2018 Application of different clustering approaches to hydro-climatological catchment regionalization in mountainous regions, a case study in Utah State. *Journal of Mountain Science* **15** (3), 461–484.

Skirvin, S., Kidwell, M., Biedenbender, S., Henley, J. P., King, D., Collins, C. H., Moran, S. & Weltz, M. 2008 Vegetation data, walnut gulch experimental watershed, Arizona, United States. *Water Resources Research* **44**, W05S08.

Srinivas, V. V., Tripathi, S., Rao, A. R. & Govindaraju, R. S. 2008 Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering. *Journal of Hydrology* **348**, 148–166.

Stone, J., Nichols, M., Goodrich, D. & Buono, J. 2008 Long-term runoff database, walnut gulch experimental watershed, Arizona, United States. *Water Resources Research* **44**, W05S05.

Suhr, D. 2006 *Exploratory or Confirmatory Factor Analysis*. SAS Users Group International Conference, SAS Institute, Cary, NC, pp. 1–17.

Tian, F., Hu, H., Lei, Z. & Sivapalan, M. 2006 Extension of the representative elementary watershed approach for cold regions via explicit treatment of energy related processes. *Hydrology and Earth System Science* **10**, 619–644.

USDA, United States Department of Agriculture 2007 Agricultural Research Service, Southwest Watershed Research Center, Tucson, Arizona. http://www.tucson.ars.ag.gov/swrc_site/research/wgew.

Ward Jr, J. H. 1963 Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58**, 236–244.

Wehrens, R. & Buydens, L. M. C. 2007 Self- and super-organizing maps in R: the kohonen package. *Journal of Statistical Software* **21** (5), 1–19.

Weiss, A. 2001 Topographic Position and Landforms Analysis. In: *ESRI User Conference*, San Diego, CA.

White, D., Richman, M. & Yarnal, B. 1991 Climate regionalization and rotation of principal components. *International Journal of Climatology* **11**, 1–25.

Wolock, D. M., Winter, T. C. & McMahon, G. 2004 Delineation and evaluation of hydrologic-landscape regions in the United States using geographic information system tools and multivariate statistical analyses. *Environmental Management* **34**, S71–S88.

Wood, E. F., Sivapalan, M., Beven, K. & Band, L. 1988 Effects of spatial variability and scale with implications to hydrologic modeling. *Journal of Hydrology* **102** (1–4), 29–47.

Xie, X. L. & Beni, G. 1991 A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13** (8), 841–847.

Zhang, Y., Wang, W., Zhang, X. & Li, Y. 2008 A cluster validity index for fuzzy clustering. *Information Sciences* **178**, 1205–1218.

First received 7 January 2018; accepted in revised form 31 July 2018. Available online 5 September 2018