A fuzzy hybrid clustering method for identifying hydrologic homogeneous regions
S. Saeid Mousavi Nadoushani, Naser Dehghanian and Bahram Saghaian

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)

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
REAA Representative Elementary Area
REW Representative Elementary Watersheds
SOFM Self-Organizing Features Maps
TPI Topographic Position Index
VAR Variable
WGEW Walnut Gulch Experimental Watershed

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. 2005) 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:

\[
x_{jn} = \frac{y_{jn} - y_{n\text{min}}}{y_{n\text{max}} - y_{n\text{min}}} \tag{1}
\]

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

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

which is subject to the following constraints:

\[
J = \sum_{i=1}^{c} u_{ij} = 1 \quad \forall \ j \in \{1, \ldots, N\} \tag{3}
\]

\[
0 < \sum_{j=1}^{N} u_{ij} < N \quad \forall \ i \in \{1, \ldots, c\} \tag{4}
\]

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 \(j\)th rescaled feature cell belonging to the \(i\)th 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 \([1, \infty)\). 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 \(j\)th layer, \(c_i\) is the center of the \(i\)th 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:

\[
u_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{d_{ij}}{d_{ik}} \right)^m} \quad \text{for } i = 1, \ldots, c \text{ & } k = 1, \ldots, N \tag{5}
\]

\[
c_i = \frac{\sum_{j=1}^{N} u_{ij}^m x_j}{\sum_{j=1}^{N} u_{ij}^m} \quad \text{for } i = 1, \ldots, c \tag{6}
\]

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 \(\alpha\)-cut \((\alpha_{\text{low}}, \alpha_{\text{high}})\) for the degree of membership based on the number of clusters \((c)\). If \(\alpha_{\text{high}} \geq 1 - 1/c\), then that cell definitely belongs to the cluster, and if \(\alpha_{\text{low}} < 1/c\), 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:

\[
CI_i = 1 - (u_{i,1\text{st max}} - u_{i,2\text{nd max}}) \cdot CI \in [0,1] \tag{7}
\]

where \(u_{i,1\text{st max}}\) and \(u_{i,2\text{nd max}}\) 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 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)

<table>
<thead>
<tr>
<th>Index</th>
<th>Source</th>
<th>Equation</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partition coefficient</td>
<td>Bezdek (1974a, 1974b)</td>
<td>( V_{PC} = \frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{c} u_{ij}^2 )</td>
<td>( V_{PC} ): Partition coefficient</td>
</tr>
<tr>
<td>Partition entropy</td>
<td>Bezdek (1974a, 1974b)</td>
<td>( V_{PE} = -\frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{c} u_{ij} \log(u_{ij}) )</td>
<td>( V_{PE} ): Partition entropy</td>
</tr>
<tr>
<td>Fukuyama–Sugeno</td>
<td>Fukuyama &amp; Sugeno (1989)</td>
<td>( V_{FS} = \frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{c} u_{ij}^2 \left(</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \bar{c} = \frac{1}{N} \sum_{j=1}^{N} x_j )</td>
<td>( V_{EXB} ): The extended FCM</td>
</tr>
<tr>
<td>Extended Xie–Beni</td>
<td>Xie &amp; Beni (1991) and Pal &amp; Bezdek (1995)</td>
<td>( V_{EXB} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} u_{ij}^2</td>
<td></td>
</tr>
<tr>
<td>Kwon’s index</td>
<td>Kwon (1998)</td>
<td>( V_k = \frac{\sum_{i=1}^{N} \sum_{j=1}^{c} u_{ij}^2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \bar{c} = \frac{1}{N} \sum_{j=1}^{N} x_j )</td>
<td>( c ): Number of clusters</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( u_{ij} ): Degree of membership</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( x_j ): Rescaled feature value for each cell</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( c_i ) and ( c_j ): The fuzzy centroids of the ith cluster</td>
</tr>
</tbody>
</table>
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 \( c \):

\[
R^2 = \frac{\sum (c_i - \bar{x}_i)^2}{\sum (x_i - \bar{x}_i)^2}
\]  

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

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 × 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.

---

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

<table>
<thead>
<tr>
<th>No.</th>
<th>Input data</th>
<th>Unit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DEM</td>
<td>m</td>
<td>Digital elevation model (DEM) with 150 m cell size</td>
</tr>
<tr>
<td>2</td>
<td>Normalized difference vegetation index (NDVI)</td>
<td>–</td>
<td>Constructed based on LANDSAT bands 3 and 4 (red and near-infrared)</td>
</tr>
<tr>
<td>3</td>
<td>Soil saturated hydraulic conductivity (Ksat)</td>
<td>mm/hr</td>
<td>Available on WGEW-USDA website prepared in four categories with a range of 8.36</td>
</tr>
<tr>
<td>4</td>
<td>Topographic position index (TPI)</td>
<td>m</td>
<td>TPI reflects the tendency of each cell to saturate.</td>
</tr>
<tr>
<td>5</td>
<td>Slope aspect</td>
<td>–</td>
<td>In 9 categories (1 flat area and 8 other main directions)</td>
</tr>
<tr>
<td>6</td>
<td>Plan curvature</td>
<td>–</td>
<td>Influences the convergence/divergence of surface runoff</td>
</tr>
<tr>
<td>7</td>
<td>Profile curvature</td>
<td>–</td>
<td>Influences the runoff acceleration/deceleration</td>
</tr>
<tr>
<td>8</td>
<td>Slope</td>
<td>Percent</td>
<td>Slope (in percent)</td>
</tr>
<tr>
<td>9</td>
<td>Curve number (CN)</td>
<td>–</td>
<td>Based on the combination of soil and land use</td>
</tr>
<tr>
<td>10</td>
<td>Distance from the nearest stream (Distance)</td>
<td>m</td>
<td>An important factor in flow transfer to the outlet</td>
</tr>
</tbody>
</table>
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).
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, \( Euc\_\text{mean} \) 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 \( N\_\text{Umax} = 0 \) 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 \( (N\_\text{Umax} = 0) \), 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

<table>
<thead>
<tr>
<th>Component (VAR)</th>
<th>Initial eigenvalues</th>
<th>% of variance</th>
<th>Cumulative variance %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.253</td>
<td>82.529</td>
<td>82.529</td>
</tr>
<tr>
<td>2</td>
<td>1.108</td>
<td>11.082</td>
<td>93.611</td>
</tr>
<tr>
<td>3</td>
<td>0.488</td>
<td>4.876</td>
<td>98.487</td>
</tr>
<tr>
<td>4</td>
<td>0.150</td>
<td>1.505</td>
<td>99.991</td>
</tr>
<tr>
<td>5</td>
<td>0.001</td>
<td>0.009</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>1.392 \times 10^{-5}</td>
<td>0.000</td>
<td>100</td>
</tr>
<tr>
<td>7</td>
<td>6.275 \times 10^{-6}</td>
<td>6.275 \times 10^{-5}</td>
<td>100</td>
</tr>
<tr>
<td>8</td>
<td>9.385 \times 10^{-7}</td>
<td>9.385 \times 10^{-6}</td>
<td>100</td>
</tr>
<tr>
<td>9</td>
<td>8.995 \times 10^{-8}</td>
<td>8.995 \times 10^{-7}</td>
<td>100</td>
</tr>
<tr>
<td>10</td>
<td>1.128 \times 10^{-8}</td>
<td>1.128 \times 10^{-7}</td>
<td>100</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th>Component (VAR)</th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>1</td>
</tr>
<tr>
<td>VAR1 (NDVI)</td>
<td>0.573</td>
</tr>
<tr>
<td>VAR2 (Ksat)</td>
<td>0.573</td>
</tr>
<tr>
<td>VAR3 (TPI)</td>
<td>0.901</td>
</tr>
<tr>
<td>VAR4 (Aspect)</td>
<td>0.901</td>
</tr>
<tr>
<td>VAR5 (Curvature-plan)</td>
<td>0.901</td>
</tr>
<tr>
<td>VAR6 (Curvature-profile)</td>
<td>0.901</td>
</tr>
<tr>
<td>VAR7 (Slope)</td>
<td>0.901</td>
</tr>
<tr>
<td>VAR8 (CN)</td>
<td>-0.053</td>
</tr>
<tr>
<td>VAR9 (Distance)</td>
<td>0.433</td>
</tr>
<tr>
<td>VAR10 (DEM)</td>
<td>0.901</td>
</tr>
</tbody>
</table>
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 5 | Best SOFM architecture

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

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

<table>
<thead>
<tr>
<th>Clustering scenario</th>
<th>Index</th>
<th>$V_{EXB}$</th>
<th>$V_K$</th>
<th>$V_{FS}$</th>
<th>$OF$</th>
<th>Technique I $V_{PC}$</th>
<th>$V_{PE}$</th>
<th>Technique II Cl_mean</th>
<th>Euc_mean</th>
<th>Technique III $N_Umax$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Four-component</td>
<td>TPI-Distance-CN-P25</td>
<td>min (1.1 &lt; $m$ &lt; 2.2)</td>
<td>0.495</td>
<td>3,088</td>
<td>–21,395</td>
<td>1,295</td>
<td>0.205</td>
<td>0.083</td>
<td>0.058</td>
<td>1.762</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>max (1.1 &lt; $m$ &lt; 2.2)</td>
<td>1.785</td>
<td>11,110</td>
<td>–7,544</td>
<td>14,515</td>
<td>0.951</td>
<td>1.723</td>
<td>0.834</td>
<td>1.936</td>
<td>5,850</td>
</tr>
<tr>
<td></td>
<td></td>
<td>for $c = 6$ and $m = 1.6$</td>
<td>0.628</td>
<td>5,914</td>
<td>–10,376</td>
<td>6,586</td>
<td>0.574</td>
<td>0.874</td>
<td>0.460</td>
<td>1.886</td>
<td>2,061</td>
</tr>
<tr>
<td>2. Four-component</td>
<td>TPI-Distance-CN-P50</td>
<td>min (1.1 &lt; $m$ &lt; 2.6)</td>
<td>0.498</td>
<td>3,099</td>
<td>–19,217</td>
<td>1,296</td>
<td>0.188</td>
<td>0.075</td>
<td>0.056</td>
<td>1.741</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>max (1.1 &lt; $m$ &lt; 2.6)</td>
<td>2.339</td>
<td>14,544</td>
<td>–1,096</td>
<td>12,191</td>
<td>0.957</td>
<td>2.044</td>
<td>0.907</td>
<td>1.936</td>
<td>5,866</td>
</tr>
<tr>
<td></td>
<td></td>
<td>for $c = 6$ and $m = 1.6$</td>
<td>0.618</td>
<td>5,851</td>
<td>–10,109</td>
<td>6,619</td>
<td>0.563</td>
<td>0.892</td>
<td>0.477</td>
<td>1.882</td>
<td>1,976</td>
</tr>
<tr>
<td>3. Four-component</td>
<td>TPI-Distance-CN-P100</td>
<td>min (1.1 &lt; $m$ &lt; 2.3)</td>
<td>0.503</td>
<td>3,127</td>
<td>–21,166</td>
<td>1,036</td>
<td>0.284</td>
<td>0.099</td>
<td>0.062</td>
<td>1.762</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>max (1.1 &lt; $m$ &lt; 2.3)</td>
<td>2.569</td>
<td>16,170</td>
<td>–7,926</td>
<td>12,184</td>
<td>0.943</td>
<td>1.757</td>
<td>0.871</td>
<td>1.944</td>
<td>5,861</td>
</tr>
<tr>
<td></td>
<td></td>
<td>for $c = 6$ and $m = 1.6$</td>
<td>0.691</td>
<td>4,306</td>
<td>–10,395</td>
<td>6,649</td>
<td>0.563</td>
<td>0.888</td>
<td>0.487</td>
<td>1.902</td>
<td>1,915</td>
</tr>
<tr>
<td>4. Three-component</td>
<td>TPI-CN-P50</td>
<td>min (1.1 &lt; $m$ &lt; 2.7)</td>
<td>0.171</td>
<td>1,166</td>
<td>–17,259</td>
<td>280</td>
<td>0.257</td>
<td>0.055</td>
<td>0.027</td>
<td>1.516</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>max (1.1 &lt; $m$ &lt; 2.7)</td>
<td>0.659</td>
<td>4,154</td>
<td>–2,752</td>
<td>7,411</td>
<td>0.968</td>
<td>1.895</td>
<td>0.812</td>
<td>1.670</td>
<td>6,456</td>
</tr>
<tr>
<td></td>
<td></td>
<td>for $c = 6$ and $m = 1.6$</td>
<td>0.386</td>
<td>2,416</td>
<td>–10,509</td>
<td>3,896</td>
<td>0.690</td>
<td>0.642</td>
<td>0.326</td>
<td>1.656</td>
<td>3,261</td>
</tr>
<tr>
<td>Previous studies</td>
<td></td>
<td>min</td>
<td>0.15</td>
<td>–602.51</td>
<td>81.87</td>
<td>160</td>
<td>0.99</td>
<td>0.05</td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>max</td>
<td>14.49</td>
<td>852.36</td>
<td>3,549.6</td>
<td>1,280</td>
<td>0.21</td>
<td>2.11</td>
<td>1</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
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 ($N_{U_{\text{max}}}$) belonging to the clusters with a certain degree of membership (as $\alpha_{\text{high}}$ decreases) has a decreasing trend with increasing $m$ and $c$.
- $N_{U_{\text{max}}}$ for clustering scenario 4 was greater than the ones in other clustering scenarios for all values of m.

Figure 5 | Variations of selected fuzzy validation indices for clustering scenario 4 (TP-CN-P50): (a) $V_{PC}$, (b) $V_{PE}$, (c) $V_{FS}$, (d) $V_K$, and (e) $V_{EXL}$.
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 ≤ c ≤ 11). In most cases of c, the point of VPC and VPE 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>
<thead>
<tr>
<th>Clustering scenario (input components)</th>
<th>Range of m</th>
<th>Optimum fuzzification parameter (m)</th>
<th>Optimum number of cluster (c)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Technique I</td>
<td>Technique II</td>
</tr>
<tr>
<td>Scenario 1 (TPI-CN-Distance-P50)</td>
<td>1.1 ≤ m ≤ 2.2</td>
<td>1.6</td>
<td>7</td>
</tr>
<tr>
<td>Scenario 2 (TPI-CN-Distance-P25)</td>
<td>1.1 ≤ m ≤ 2.6</td>
<td>1.6</td>
<td>8</td>
</tr>
<tr>
<td>Scenario 3 (TPI-CN-Distance-P100)</td>
<td>1.1 ≤ m ≤ 2.3</td>
<td>1.6</td>
<td>4</td>
</tr>
<tr>
<td>Scenario 4 (TPI-CN-P50)</td>
<td>1.1 ≤ m ≤ 2.7</td>
<td>1.7</td>
<td>6</td>
</tr>
</tbody>
</table>
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.

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

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

where $f$ is the number of selected components (in this study, $f = 3$ for three-component and 4 for four-component clustering), $x_{if}$ and $\bar{c}_{if}$ 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 centers.
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 km²) and 4% (6 km²) 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 × 0.5 cm = 0.25 cm² on the map. With scale 1:24,000 used in this study, the threshold area is 120 m × 120 m = 14,400 (m²), almost 1.44% of the total basin area (2.16 km²). 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 \( \alpha_{\text{High}} = (1-1/6) = 0.833 \). These results indicate that the three-component clustering map is more reliable and accurate than the four-component map.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Number of cells</th>
<th>Area (%)</th>
<th>Mean CI</th>
<th>Mean Umax</th>
<th>Mean TPI</th>
<th>Mean CN</th>
<th>Mean Distance (m)</th>
<th>Mean P50 (mm)</th>
<th>MARD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>276</td>
<td>4.44</td>
<td>0.400</td>
<td>0.706</td>
<td>3.53</td>
<td>77.99</td>
<td>1,223.39</td>
<td>68.13</td>
<td>1.30</td>
</tr>
<tr>
<td>2</td>
<td>891</td>
<td>14.83</td>
<td>0.410</td>
<td>0.708</td>
<td>–0.24</td>
<td>87.37</td>
<td>482.49</td>
<td>65.45</td>
<td>3.50</td>
</tr>
<tr>
<td>3</td>
<td>849</td>
<td>18.99</td>
<td>0.569</td>
<td>0.609</td>
<td>0.26</td>
<td>72.56</td>
<td>1,358.51</td>
<td>68.58</td>
<td>2.81</td>
</tr>
<tr>
<td>4</td>
<td>1,227</td>
<td>24.72</td>
<td>0.422</td>
<td>0.729</td>
<td>–0.18</td>
<td>59.11</td>
<td>473.41</td>
<td>65.15</td>
<td>3.20</td>
</tr>
<tr>
<td>5</td>
<td>2,090</td>
<td>20.89</td>
<td>0.449</td>
<td>0.715</td>
<td>–0.38</td>
<td>60.38</td>
<td>249.90</td>
<td>70.48</td>
<td>2.21</td>
</tr>
<tr>
<td>6</td>
<td>921</td>
<td>16.14</td>
<td>0.614</td>
<td>0.595</td>
<td>0.036</td>
<td>63.14</td>
<td>797.02</td>
<td>74.08</td>
<td>3.88</td>
</tr>
<tr>
<td>Total</td>
<td>6,210</td>
<td>100</td>
<td>0.477</td>
<td>0.684</td>
<td>0.504</td>
<td>70.09</td>
<td>764.12</td>
<td>68.64</td>
<td>2.81</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Number of cells</th>
<th>Area (%)</th>
<th>Mean CI</th>
<th>Mean Umax</th>
<th>Mean TPI</th>
<th>Mean CN</th>
<th>Mean Distance (m)</th>
<th>Mean P50 (mm)</th>
<th>MARD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,243</td>
<td>19.97</td>
<td>0.398</td>
<td>0.731</td>
<td>–0.063</td>
<td>74.81</td>
<td>70.74</td>
<td></td>
<td>2.28</td>
</tr>
<tr>
<td>2</td>
<td>821</td>
<td>13.19</td>
<td>0.399</td>
<td>0.755</td>
<td>–0.281</td>
<td>59.15</td>
<td>74.86</td>
<td></td>
<td>1.20</td>
</tr>
<tr>
<td>3</td>
<td>1,950</td>
<td>31.00</td>
<td>0.274</td>
<td>0.829</td>
<td>–0.165</td>
<td>57.93</td>
<td>69.66</td>
<td></td>
<td>0.39</td>
</tr>
<tr>
<td>4</td>
<td>1,064</td>
<td>17.09</td>
<td>0.243</td>
<td>0.848</td>
<td>–0.210</td>
<td>59.01</td>
<td>64.48</td>
<td></td>
<td>5.12</td>
</tr>
<tr>
<td>5</td>
<td>918</td>
<td>14.75</td>
<td>0.366</td>
<td>0.749</td>
<td>–0.168</td>
<td>87.10</td>
<td>64.76</td>
<td></td>
<td>8.49</td>
</tr>
<tr>
<td>6</td>
<td>249</td>
<td>4.00</td>
<td>0.337</td>
<td>0.746</td>
<td>3.472</td>
<td>77.97</td>
<td>68.34</td>
<td></td>
<td>1.31</td>
</tr>
<tr>
<td>Total</td>
<td>6,225</td>
<td>100</td>
<td>0.326</td>
<td>0.788</td>
<td>0.451</td>
<td>69.33</td>
<td>68.81</td>
<td></td>
<td>3.13</td>
</tr>
</tbody>
</table>
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 $\alpha$-cut is required. In this study, the criterion of $(\alpha$-cut $= 1 - 1/c = 0.833$, for $c = 6$) was applied. Although $\alpha$-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 ($CI_e(0.2–0.4)$).

- Application of the selected conventional validation indices ($V_{PC}$, $V_{PE}$, $V_{IS}$, $V_{EEX}$, 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


