A hybrid regionalization method based on canonical correlation analysis and cluster analysis: a case study in northern Iran

Ali Ahani, S. Saeid Mousavi Nadoushani and Ali Moridi

ABSTRACT

The performance of regionalization methods used for regional flood frequency analysis is affected considerably by the features used to identify the homogeneous regions (e.g., climatological, meteorological, geomorphological, and physiographic characteristics of the watersheds). In this study, a regionalization method is proposed that takes advantage of the two widely used techniques in regionalization of watersheds: canonical correlation analysis and cluster analysis. In the proposed method, the canonical correlation analysis is utilized to select or weight features that then will be used by a hybrid clustering algorithm for regionalization of watersheds. The proposed method is applied to Sefidrud basin, located in the north of Iran, to implement regionalization with two, three, four, and five regions. Performance assessment of the proposed method shows that all the options of the proposed method can be effective alternatives to some common regionalization methods to improve the homogeneity of the regions. The results indicate that the method can satisfy the homogeneity conditions approximately for all the regions which were identified in the study area.

Key words | canonical correlation analysis, cluster analysis, feature selection, feature weighting, regionalization

NOTATION

| Term   | Definition                                                                 |
|--------|-----------------------------------------------------------------------------|
| FFA    | Flood frequency analysis                                                    |
| RFFA   | Regional flood frequency analysis                                           |
| CCA    | Canonical correlation analysis                                              |
| WAKM   | A hybrid clustering algorithm consisting of Ward’s algorithm and K-means   |
| ASW    | Average silhouette width                                                    |
| CCA-WAKM | A set of four implementation options of the proposed regionalization method including the combination of CCA and WAKM in which the feature vectors consisting of the canonical variables of the watershed features are used in clustering by WAKM |
| CCA-WAKM\textsubscript{1} | An option for implementing CCA-WAKM in which the values of the first canonical variable of the watershed features are used as the feature vectors of sites |
| CCA-WAKM\textsubscript{1,2} | An option for implementing CCA-WAKM in which the values of the first and second canonical variables of the watershed features are used as the features of the feature vectors of sites |
| CCA-WAKM\textsubscript{1,3} | An option for implementing CCA-WAKM in which the values of the first and third canonical variables of the watershed features are used as the features of the feature vectors of sites |

doi: 10.2166/nh.2019.105
CCA-WAKM\textsubscript{1,2,3} An option for implementing CCA-WAKM in which the values of the first, second and third canonical variables of the watershed features are used as the features of the feature vectors of sites

CCA-wWAKM An implementation option of the proposed regionalization method in which the coefficients of the watershed features in the first canonical variables of the watershed features are used as the weights of the features in clustering by WAKM

**INTRODUCTION**

Flood frequency analysis (FFA) is used to estimate the magnitude of a flood with a specified return period or estimate the return period of a flood with a specified magnitude. Flood quantiles can be estimated by at-site FFA using only flood data recorded at the site of interest. However, in many cases, the length of flood data records in sites of interest are not appropriate to provide reliable flood estimates. In such situations, regional flood frequency analysis (RFFA) is an efficient approach to compensate for the temporal shortage of flood data records by pooling flood data over a number of sites with similar flood generation mechanisms.

The objective of the regionalization is to identify homogeneous regions, i.e., groups of sites with similar flood generation mechanisms (Hosking & Wallis 1997). In a homogeneous region, the flood frequency distribution varies from site to site with only a site-specific factor named the index-flood. RFFA based on index-flood was first introduced by Dalrymple (1960).

The identification of homogeneous regions for RFFA is required to find an appropriate regional flood frequency distribution. However, identifying a group of sites satisfying homogeneity conditions is not always possible easily. When the regionalization features not directly related to flood data records (e.g., climatological, meteorological, geomorphological, and physiographic characteristics of the watersheds), it is difficult to assign all the watersheds to the regions where all of them satisfy homogeneity conditions.

One of the most widely used methods for regionalization is cluster analysis. The cluster analysis methods are multivariate statistical analysis methods which have been utilized by many researchers in several hydrological studies especially for the regionalization (e.g., Acreman & Sinclair 1986; Burn 1989; Hall & Minns 1999; Jingyi & Hall 2004; Lin & Chen 2006; Ramachandra Rao & Srinivas 2006a, 2006b; Srinivas et al. 2008; Chen & Hong 2012; Toth 2013). While in traditional methods of regionalization, regions were identified based on administrative borders or geographical contiguity, in the methods like cluster analysis, different types of features effective on flood generation mechanism, such as physiographic features or meteorological attributes, can be used as regionalization features. By applying the traditional methods, it is very difficult to identify homogeneous regions, because geographical contiguity of sites does not result in a similarity in their flood generation mechanism. On the other hand, in cluster analysis methods, the regions may be identified based on similarity of sites in terms of various features such as physiographic attributes, meteorological characteristics, plant cover, land use, etc. Thus, in the new methods, regions are not identified essentially based on geographical contiguity.

Region of influence (ROI) is another widely used approach for RFFA that was developed by Burn (1990). In ROI, a region of influence, which is a hydrologically homogeneous neighborhood, is formed for each watershed in a study area. ROI has been used in several regional frequency analysis studies and its performance evaluated in different case studies (e.g., Zrinji & Burn 1994; Burn 1997; Castellarin et al. 2001).

Clustering algorithms can be divided into hierarchical and partitional clustering algorithms (Ramachandra Rao & Srinivas 2008). Hierarchical algorithms include agglomerative algorithms and divisive algorithms where the agglomerative hierarchical algorithms have been used for regionalization of watersheds in several RFFA studies (e.g., Mosley 1981; Tasker 1982; Acreman & Sinclair 1986; Nathan & McMahon 1990; Burn et al. 1997; Hosking & Wallis 1997; Ramachandra Rao & Srinivas 2006a). One of the most important advantages of hierarchical algorithms is that they often do not require the determination of initial conditions (such as the determination of initial cluster centers). On the other hand, a noticeable limitation of hierarchical
algorithms is that after assigning a data point to a cluster, it is not possible to move it between clusters. Partitional algorithms, which often are based on the minimization of an objective function, require the determination of initial conditions, such as the initial cluster centers, but these algorithms often provide the benefits of the possibility of moving data points between different clusters in different iterations of the algorithm. One of the most widely used partitional clustering algorithms in regional frequency analysis studies is the K-means algorithm (e.g., Wiltshire 1986; Burn 1989; Bhaskar & O’Connor 1989; Burn & Goel 2000; Ramachandra Rao & Srinivas 2006a; Jin et al. 2017; Xie et al. 2018).

Ramachandra Rao & Srinivas (2006a) investigated the performances of combinations of the three hierarchical algorithms with one partitional algorithm for regionalization of 245 watersheds in Indiana State, USA. They used single linkage, complete linkage, and Ward’s algorithm as hierarchical algorithms to specify initial cluster centers for K-means as a partitional algorithm. The quality of the clusters formed by each of the proposed hybrid algorithms was evaluated according to values of four cluster validity indices: cophenetic correlation coefficient (CPCC), silhouette width, Dun’s index, and Davies–Bouldin index. Also, the homogeneity of the regions identified by each algorithm was assessed based on the values of the heterogeneity measures proposed by Hosking & Wallis (1993). The proposed hybrid algorithms showed better performances in comparison with the hierarchical and partitional clustering algorithms. In addition, among the proposed algorithms, the combinations of Ward’s and K-means algorithms (WAKM) provided the best regionalization results for RFRA in the study area.

From another aspect, clustering methods can be divided into hard and fuzzy clustering. In hard clustering, each data point belongs to only one cluster and does not belong to any other cluster. On the other hand, in fuzzy clustering, each data point can be assigned to all clusters simultaneously with specified degrees of membership between 0 and 1. The sum of the degrees of membership of each data point in all clusters is equal to 1. The most popular fuzzy clustering algorithm used in regional frequency analysis studies to implement regionalization is the fuzzy C-means (FCM) clustering algorithm (e.g., Hall & Minns 1999; Ramachandra Rao & Srinivas 2006b; Srinivasa Raju & Nagesh Kumar 2008; Sadri & Burn 2011; Asong et al. 2015; Basu & Srinivas 2015). In some studies, combinations of hard and fuzzy clustering algorithms have been studied for regionalization. Srinivas et al. (2008) used a combination of self-organizing maps (SOM) and FCM to implement regionalization of Indiana State watersheds. The results showed that the method was effective in forming homogeneous regions. Farsadnia et al. (2014) also carried out a similar study for regionalization of the watersheds of Mazandaran province in northern Iran and obtained similar results. Ahani & Mousavi Nadoushani (2016) studied the fuzzy development of the hybrid clustering algorithms proposed by Ramachandra Rao & Srinivas (2006a) by combining single linkage, complete linkage, average linkage hierarchical algorithms, and Ward’s algorithm, as well as SOM with C-means fuzzy clustering algorithm for performing regionalization of the watersheds in the Sefidrud watershed. The study of the size of the formed regions, the values of the clustering validity indices, and also the $SH$ heterogeneity indices showed that, in general, the combination of the Ward and SOM algorithms with FCM algorithm provide the best results for regionalization of the studied region in order to analyze the flood regional frequency.

The ability of cluster analysis methods in dealing with multivariate analysis problems and reducing the need for visual judgments and time-consuming assessments are the benefits of these methods for regionalization of watersheds. However, there are some issues that may affect the efficiency of cluster analysis methods for regionalization of watersheds. In regionalization by cluster analysis methods, each watershed is represented by a vector that includes values of a set of features affecting flood generation mechanism. The feature vectors are used to evaluate similarity of the watersheds. The identification of the feature vectors to be used in clustering is one of the most challenging issues in regionalization studies (e.g., Nezhad et al. 2010; Di Prinzio et al. 2011; Razavi & Coulibaly 2015; Ahani et al. 2018).

One of the useful methods to identify and select the effective features on the flood generation mechanism of watersheds is canonical correlation analysis (CCA). CCA (Hotelling 1936) is a method for describing the correlation between two sets of variables (Cavadias 1990). Cavadias (1990) developed a method based on CCA to determine the hydrological neighborhoods and estimate flood quantile. Also, Cavadias et al. (2001) proposed a method based on
the use of CCA in order to determine homogeneous regions or hydrological neighborhoods for flood estimation in both gauged and ungauged sites. The proposed method was useful to identify effective watershed features on flood generation mechanism. However, the features useful for identification of homogeneous regions were selected based on visual judgments on the similarities between patterns of data points in original feature space and canonical space. The application of CCA in RFFA was studied by several researchers (e.g., Ribeiro-Correa et al. 1995; GREHYS 1996a, 1996b; Ouarda et al. 2001, 2008) and the results indicated the desirable effects of CCA on RFFA.

In a method introduced by Ilorme & Griffis (2013), CCA was initially used along with some other multivariate analysis methods to identify the watershed features influencing the flood generation mechanism. Then, the selected features were used to perform regionalization by Ward’s clustering algorithm. The method reduced the need for visual judgment to identify homogeneous regions and select regionalization features. Also, it overcame the visual judgment limitation. However, the different effects of the different features on the final regionalization was not considered in the proposed method. In addition, skipping a number of features with relatively lower values of correlation coefficient might significantly reduce homogeneity of regions and accuracy of flood quantile estimation (Basu & Srinivas 2014).

In general, determining the relationship between the watershed features (such as geographical location characteristics, physiographic attributes, geological features, land-use, plant cover, etc.) and the flood-related features (such as flood statistics) can be considered as an important advantage of CCA-based RFFA methods. However, most CCA-based regionalization methods depend on visual judgments to some extent, and in some cases, they are theoretically limited to two-dimensional space.

The main objective of the current study is to propose an efficient regionalization method focusing on feature selection and feature weighting to improve the homogeneity of the regions. To this aim, a new hybrid method is proposed by combining CCA and cluster analysis in order to take the advantages of both of them and overcome their limitations in regionalization of watersheds for RFFA. After describing the proposed method, some implementation options of the method are presented for regionalization of watersheds in Sefidrud basin located in the north of Iran. Then the performance of implementation options of the method is compared with that of a common regionalization method in the study area.

**STUDY AREA AND DATA**

Sefidrud basin in the north of Iran with a total area about 59,200 km² was chosen as a study area to evaluate the performance of proposed methods for regionalization of watersheds. Sefidrud River is formed in the confluence of two rivers, named Shahrud and Ghezel-Ozan, and flows into the Caspian Sea. Thirty-nine gauged sites with unregulated flow in Sefidrud basin were selected for this study and their watershed features were extracted for regionalization (Figure 1). The annual maximum flood data records of the sites of interest were obtained from the database of the Iran Water Resources Management Company. The flood data records in the selected sites cover the time period from 1967 to 2012 and the average length of flood data records is about 23 years. The total number of flood data is equal to 898 station-years and record length in the sites varies between 10 and 39 years.

Longitude, latitude, elevation from the sea level, drainage area, mean annual precipitation, and the runoff coefficient were selected as the watershed features contributing to the regionalization procedure (Table 1). It is worth noting that four sites have runoff coefficients greater than 1 because the watersheds of these sites are located in areas with karst geologic structures.

The features were selected based on the availability of the relevant data and their potential role in the flood generation mechanism. The longitude, latitude, and elevation from the sea level were selected, because of the special geographical situation of the study area. In fact, the noticeable variability in values of longitude, latitude, and elevation from the sea level in the study area may affect climatological and meteorological conditions of the sites considerably. The considerable variation of these features in the study area can affect the flood generation mechanisms of the watersheds noticeably. The precipitation and the drainage area are considered for regionalization in several RFFA studies due to their pivotal role in flood generation (e.g., Ramachandra...
Precipitation is often the main factor generating floods (Ramachandra Rao & Srinivas 2006a; 2006b; Srinivas et al. 2008; Farsadnia et al. 2014) and so, mean annual precipitation of watersheds was selected as one of the watershed features. Additionally, in hydrological models, the drainage area often is considered as one of the most important factors in estimating the flood magnitudes (Hosking & Wallis 1997; Ramachandra Rao & Srinivas 2008). Thus, it is logical to use the drainage area as one of the watershed features in the regionalization procedure. Also, the runoff coefficient was selected as it determines the ratio of precipitation transformed to runoff and, therefore, it may be useful to identify homogeneous regions (e.g., Ramachandra Rao & Srinivas 2006a, 2006b; Srinivas et al. 2008; Basu & Srinivas 2015).

In order to modify the asymmetry of drainage area values, the logarithmic transformation was applied to them. When there is a considerable asymmetry in the drainage area values, a few sites in the tail of the distribution may form small groups or regions (in terms of station-years) because they are completely separated from other sites in a study area. Such small regions are not desirable for RFFA because it is not possible to provide reliable flood estimates for long return periods in small regions (Hosking & Wallis 1997; Basu & Srinivas 2014). Also, the values of all the features were standardized by Equation (1) in order to

| Feature                  | Range       | Mean   | Standard deviation |
|--------------------------|-------------|--------|--------------------|
| Longitude (dd)           | 47.05-51.07 | 48.74  | 1.30               |
| Latitude (dd)            | 35.18-37.53 | 36.55  | 0.73               |
| Elevation (m a.s.l.)     | 40-2,800    | 1,376.22 | 649.20            |
| Drainage area (km²)      | 29-49,300   | 5,591.38 | 11,569.86        |
| Mean annual precipitation (mm) | 184-1,400 | 467.70  | 323.62            |
| Runoff coefficient       | 0.03-1.32   | 0.49   | 0.34               |

Figure 1 | Location of Sefidrud Basin and the hydrometric stations.
eliminate the effects of the differences in dimensions and variances of the different features:

\[ y_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j} \]  

(1)

where \( x_{ij} \) is the value of the feature \( j \) in the data point \( i \); \( \bar{x}_j \) is the mean value of the feature \( j \) over the dataset, and \( S_j \) is the standard deviation of the feature \( j \) over the dataset. In addition, \( y_{ij} \) is the standardized value of the feature \( j \) for the data point \( i \).

**METHODS**

**Discordancy evaluation**

Prior to the feature selection or feature weighting steps, the flood data records were evaluated by using the discordancy measure \( D \) proposed by Hosking & Wallis (1993) in terms of the L-moments of flood data. A site is identified as discordant if \( D \) exceeds the critical value. When the number of sites is greater than 14, the critical value of \( D \) is equal to 3. This screening procedure can be performed either before regionalization for all the sites as one group, or after regionalization for the sites belonging to each region. Among the 39 sites, two sites were identified as discordant and were excluded from the regionalization process as suggested by Hosking & Wallis (1997). Therefore, the data related to the 37 remaining sites was used in the next stages of the study.

**Canonical correlation analysis**

In CCA, a canonical space is formed based on two sets of canonical variables. Each canonical variable is a linear combination of one of the two sets of original variables. If the original variables include two sets of the variables \( A = \{A_1, A_2, \ldots, A_p\} \) and \( B = \{B_1, B_2, \ldots, B_q\} \) and \( p < q \), then the two sets of the canonical variables \( V = \{V_1, V_2, \ldots, V_p\} \) and \( W = \{W_1, W_2, \ldots, W_p\} \) are formed as in Equation (2), such that the correlation between pairs of the corresponding canonical variables is maximized and the correlation between other pairs are minimized. The highest correlation is related to the canonical variables \( V_1 \) and \( W_1 \), and the lowest correlation is related to the canonical variables \( V_p \) and \( W_p \). For more details on CCA, see Hotelling (1936) and Cavadias (1990).

\[
V_i = a_{1i}A_1 + a_{2i}A_2 + \cdots + a_{pi}A_p \\
W_i = b_{1i}B_1 + b_{2i}B_2 + \cdots + b_{qi}B_q
\]  

(2)

In the present study, the six watershed features and three L-moment (Hosking 1990) ratios of flood data are used as the two sets of original variables for CCA. The three selected L-moment ratios are the linear coefficient of variation (L-CV), linear skewness (L-skewness) and linear kurtosis (L-kurtosis). They are chosen because the three \( H \) heterogeneity measures proposed by Hosking & Wallis (1997) are calculated based on the values of these three L-moment ratios. Therefore, the use of canonical variables of watershed features which are highly correlated with the canonical variables of the L-moment ratios may increase the homogeneity of the regions identified in the regionalization.

After standardization of watershed features, L-moment ratios were calculated for each site and the standardization technique was applied to the values of L-CV, L-skewness, and L-kurtosis because the standardization of both original datasets is recommended before implementing CCA (Ribeiro-Correa et al. 1995). The standardized watershed features, longitude, latitude, elevation from the sea level, drainage area, mean annual precipitation and the runoff coefficient may be represented by \( A_1, A_2, A_3, A_4, A_5, \) and \( A_6 \), respectively, hereafter. Also, \( B_1, B_2, \) and \( B_3 \) denote the standardized L-moment ratios, L-CV, L-skewness, and L-kurtosis, in this order. Then CCA was performed on the standardized dataset of the six watershed features and the standardized dataset of the three L-moment ratios. Consequently, three pairs of canonical variables were calculated. The canonical variables related to the watershed feature space are represented by \( V_1, V_2, \) and \( V_3 \), and the canonical variables related to the L-moment ratios space are denoted by \( W_1, W_2, \) and \( W_3 \).

**WAKM clustering algorithm**

Regarding the advantages of the hybrid clustering algorithms (Ramachandra Rao & Srinivas 2006a; Srinivas et al. 2006b),...
the WAKM algorithm (Ramachandra Rao & Srinivas 2006a) is used as the clustering algorithm for the regionalization of watersheds in this study. The name WAKM represents the combination of Ward’s algorithm (Ward Jr 1963) and K-means algorithm (Hartigan & Wong 1979). In this algorithm, first, by applying Ward’s algorithm to the data points, a desired number of clusters are provided. Then, the cluster centers are used as initial cluster centers for clustering the data points by the K-means algorithm (Ramachandra Rao & Srinivas 2008). More details on Ward’s and K-means algorithms are available in Ramachandra Rao & Srinivas (2008).

Cluster validity index

To compare the quality of different clusterings performed on the same dataset, cluster validity indices are used. The clustering quality is improved as the distances between the data points belonging to each cluster decrease (smaller intra-cluster distances) and the distances between the data points belonging to different clusters increase (greater inter-cluster distances) (Ramachandra Rao & Srinivas 2008).

Ramachandra Rao & Srinivas (2006a) evaluated the performances of a number of cluster validity indices to determine the optimal number of clusters in order to perform regionalization of watersheds. They concluded that the average silhouette width (ASW) is an effective measure for this purpose.

Rousseeuw (1987) defined the silhouette width for a data point $i$ in a clustered dataset, as Equation (3):

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]}$$

(3)

where $a(i)$ is the average distance of the data point $i$ from the data points with which it is placed in the same cluster; and $b(i)$ is the minimum average distance from the data point $i$ from the data points in a cluster different from the cluster that the data point $i$ belongs to. The value of $s(i)$ can be in the range $[-1, 1]$, where values close to 1 indicate the allocation of the data point $i$ to an appropriate cluster, and values close to −1 represent the assignment of the data point $i$ to an inappropriate cluster. The average silhouette width (ASW) criterion is obtained by averaging on the values of the silhouette width of all the clustered data points and hence it varies over the range $[-1, 1]$ as well (Ramachandra Rao & Srinivas 2008). Considering the acceptable performance of ASW in evaluating the quality of clusters (Ramachandra Rao & Srinivas 2006a), it was used for clustering evaluation in this study.

A hybrid method for regionalization

In the present study, a new hybrid method is proposed for regionalization of watersheds. To implement the regionalization method, four options are presented based on feature selection and one option is presented based on feature weighting. In the four feature selection-based options, after implementing CCA, the canonical variables consisting of the watershed features highly correlated with the canonical variables of the flood statistics (L-CV, L-skewness, L-kurtosis) were used as input features of the WAKM clustering algorithm. These four options are represented by CCA-WAKM$_{fi}$ in the remainder of the article. In CCA-WAKM$_{fi}$, the subscript $fi$ is an abbreviation for feature vector and can be replaced by one of the options 1, 1, 2, 1, 3, and 1, 2, 3. The four CCA-WAKM options differ in defining the feature vectors corresponding to the sites used in clustering by WAKM.

The input feature vectors of clustering for the CCA-WAKM$_1$, CCA-WAKM$_{1,2}$, CCA-WAKM$_{1,3}$, and CCA-WAKM$_{1,2,3}$ are described in Table 2. In the first option, denoted by CCA-WAKM$_1$, only the value of the first canonical variable of the watershed features is used as the feature vector of each site. In the second option, represented by CCA-WAKM$_{1,2}$, the feature vector of each site includes the values of the first and second canonical variables of the watershed features. In the third option, introduced by

| Option      | Variables of the feature vector |
|-------------|---------------------------------|
| CCA-WAKM$_1$| $V_1$                           |
| CCA-WAKM$_{1,2}$| $V_1, V_2$                     |
| CCA-WAKM$_{1,3}$| $V_1, V_3$                     |
| CCA-WAKM$_{1,2,3}$| $V_1, V_2, V_3$               |
| CCA-wWAKM   | $a_{11} A_1, a_{12} A_2, a_{13} A_3, a_{14} A_4, a_{15} A_5, a_{16} A_6$ |
CCA-WAKM<sub>1,3</sub>, the feature vector of each site contains the values of the first and third canonical variables of the watershed features. Finally, in the fourth option, denoted by CCA-WAKM<sub>1,2,3</sub>, the values of all the three canonical variables consisting of the watershed features, are used to form the feature vector of each site. Since in the space of canonical variables, the highest correlation exists between the first pair of canonical variables of the watershed features and the L-moment ratios, the first canonical variable of the watershed features is used in all the four options. Also, the second and third canonical variables of the watershed features are added to the feature vectors in different options in order to investigate their effects on the regionalization results.

In the feature weighting-based option, weights of the watershed features in the linear combination of the first canonical variable of watershed features (i.e., \( a_{11}, a_{12}, a_{13}, a_{14}, a_{15}, a_{16} \)), are used as the weights of the original watershed features (i.e., \( A_1, A_2, A_3, A_4, A_5, A_6 \)) in the regionalization by WAKM. Thus, the regionalization feature vector used this option is \([a_{11}A_1, a_{12}A_2, a_{13}A_3, a_{14}A_4, a_{15}A_5, a_{16}A_6] \). To facilitate referring to this option, the acronym CCA-wWAKM is used in the rest of the article, in which wWAKM represents weighting WAKM. The input feature vector of clustering for the CCA-wWAKM is determined in Table 2.

For implementing the feature weighting-based option CCA-wWAKM, the coefficients of the watershed features in the first canonical variable are applied to the feature vectors of the standardized watershed features as the weights. Then, WAKM is used to perform clustering based on these weighted feature vectors.

### Homogeneity assessment

Identification of homogeneous regions by regionalization methods is an important and challenging part of RFFA. Hosking & Wallis (1993) proposed three heterogeneity measures, \( H_1, H_2, \) and \( H_3 \), based on the L-moment ratios. These measures were used in several RFFA studies and were approved by the researchers (e.g., Viglione et al. 2007). For a given region, if \( H < 1 \), the region is ‘acceptably homogeneous’, if \( 1 \leq H < 2 \), the region is identified as ‘possibly heterogeneous’, and if \( H \geq 2 \), the region is regarded as ‘definitely heterogeneous’ (Hosking & Wallis 1997).

In the current study, the heterogeneity measures \( H \) are used to assess the homogeneity of the regions and a region is considered as homogeneous if \( H_1 < 1 \) and \( H_2 < 1 \) and \( H_3 < 1 \).

### RESULTS AND DISCUSSION

The coefficients of standardized watershed features and standardized L-moment ratios in the linear combinations related to the canonical variables of the watershed features and the L-moment ratios are presented in Table 3. The canonical variables of the watershed feature space are represented by \( V_1, V_2, \) and \( V_3 \), and the canonical variables of the L-moment ratios space are denoted by \( W_1, W_2, \) and \( W_3 \).

As shown in Table 4, the correlation coefficient between the first pair of canonical variables is considerably greater than the values of the correlation coefficient between the second pair of canonical variables as well as the third pair of canonical variables. The values of the correlation

### Table 3

| Standardized variable | \( V_1 \) | \( V_2 \) | \( V_3 \) | \( W_1 \) | \( W_2 \) | \( W_3 \) |
|-----------------------|---------|---------|---------|---------|---------|---------|
| \( A_1 \) (Longitude) | −0.590  | −0.021  | −1.380  | −       | −       | −       |
| \( A_2 \) (Latitude)  | −0.264  | −0.373  | 0.435   | −       | −       | −       |
| \( A_3 \) (Elevation from the sea level) | 0.272   | −1.491  | 0.412   | −       | −       | −       |
| \( A_4 \) (Drainage area) | 0.044   | −0.588  | −0.287  | −       | −       | −       |
| \( A_5 \) (Mean annual precipitation) | −0.126  | −1.579  | 0.558   | −       | −       | −       |
| \( A_6 \) (Runoff coefficient) | −0.264  | 0.614   | 0.727   | −       | −       | −       |
| \( B_1 \) (L-CV) | −       | −       | 1.004   | 0.431   | −1.644  | −       |
| \( B_2 \) (L-skewness) | −       | −       | −       | 0.034   | −0.707  | 2.827   |
| \( B_3 \) (L-kurtosis) | −       | −       | −       | −0.261  | 1.407   | −1.523  |
coefficients of the second pair of canonical variables and the third pair of canonical variables are nearly equal to each other and their difference is lower than 0.04. Therefore, it seems that the first canonical variable of the watershed features and its coefficients can play a more important role than two other canonical variables in identifying regions with better homogeneity. It is also important to note that among the coefficients of the first canonical variable of the L-moment ratios, the largest coefficient belongs to L-CV, which is the basis for calculating the heterogeneity measure $H_1$, which is more effective in identifying homogeneous regions than $H_2$ and $H_3$ heterogeneity measures according to Hosking & Wallis (1997) and Viglione et al. (2007).

In Table 5, the values of the linear correlation coefficients between the original variables and canonical variables are presented. Among watershed features, the drainage area and elevation from the sea level show the greatest positive correlations with the first canonical variables ($V_1$ and $W_1$), respectively. On the other hand, the longitude and the runoff coefficient have the largest magnitudes of negative correlations with the first canonical variables, respectively. Concerning the second canonical variables ($V_2$ and $W_2$), the highest linear correlations are those of the drainage area and the runoff coefficient, respectively, and the largest inverse correlation values are related to the elevation and the mean annual precipitation. The two features of latitude and mean annual precipitation show the highest positive correlation with the third canonical variables ($V_3$ and $W_3$), while the largest negative correlation values with these canonical variables are respectively related to the longitude and the drainage area.

All the values of correlation coefficient between the L-moment ratios and canonical variables are positive. Greatest values of correlations with first, second, and third canonical variables are related to L-CV, L-kurtosis, and L-skewness, respectively.

Since only the canonical variables of watershed features (i.e., $V_1$, $V_2$, and $V_3$) are used in the next step of the proposed regionalization method, the values of correlation coefficient between watershed features and their canonical variables are more useful to identify the watershed features that may be more effective on results of regionalization.

According to the number of selected sites in the study area and the length of their flood data records and also regarding the 5T rule (Reed et al. 1999), the regionalization was implemented by changing the number of regions from two to five. In this study, it was considered as a constraint that the smallest region (in terms of station-years) in each regionalization includes about 50 station-years flood data in order to provide flood quantiles corresponding to a ten-year return period. To evaluate the effect of the proposed methods on the homogeneity of the regions, the results of applying CCA-WAKM and CCA-wWAKM were compared with the results of applying the single WAKM clustering algorithm to feature vectors consisting of six standardized watershed features. To assess and compare the performances of the methods, they were evaluated by ASW cluster.
validity index and the heterogeneity indices $H_1$, $H_2$, and $H_3$. It should be noted that the words ‘cluster’ and ‘region’ may be used in the rest of article equivalently.

Table 6 | The values of the cluster validity index ASW for clustering by using WAKM, CCA-wWAKM, and the four CCA-WAKM options

| Number of regions | 2   | 3   | 4   | 5   |
|-------------------|-----|-----|-----|-----|
| WAKM              | 0.418 | 0.492 | 0.420 | 0.415 |
| CCA-wWAKM         | 0.557 | 0.545 | 0.484 | 0.438 |
| CCA-WAKM$_1$      | 0.655 | 0.574 | 0.557 | 0.598 |
| CCA-WAKM$_1,2$    | 0.451 | 0.516 | 0.550 | 0.430 |
| CCA-WAKM$_1,3$    | 0.459 | 0.468 | 0.525 | 0.564 |
| CCA-WAKM$_1,2,3$  | 0.367 | 0.411 | 0.431 | 0.443 |

The ASW values for implementing regionalization by each method for two, three, four, and five regions are presented in Table 6. In all cases, ASW for all the CCA-WAKM implementation options and CCA-wWAKM, are higher than those of the single WAKM. This indicates a higher quality of the final clusters resulting from the application of the proposed method in comparison with the single WAKM. The reduction of the number of dimensions or regionalization features (from six watershed features to one, two, or three canonical variables) in the CCA-WAKM implementation options compared to the WAKM may be considered as an effective factor in increasing ASW and improving the clustering quality in terms of intra-cluster compactness and inter-cluster separation. However, for CCA-wWAKM the

Figure 2 | The values of the heterogeneity measures $H_1$, $H_2$, and $H_3$ for two regions identified by WAKM, CCA-wWAKM, and CCA-WAKM.
number of regionalization features (six weighted watershed features) is equal to that of WAKM (six watershed features) and so the increase in ASW can be interpreted as an increase in the quality of clustering.

Figure 2 shows the values of the heterogeneity measures $H_1$, $H_2$, and $H_3$ for two regions identified by WAKM, CCA-wWAKM, and the four CCA-WAKM options. All the options and methods result in identifying two homogeneous regions, and only by implementing CCA-WAKM1,2, one of the regions is relatively heterogeneous based on $H_1$.

According to Figure 3, only CCA-wWAKM provides three homogeneous regions simultaneously. Both WAKM and the four CCA-WAKM options result in identifying a possibly heterogeneous region based on the values of one or two heterogeneity measures.

Figure 4 shows that the use of WAKM to identify four regions in the study area leads to identifying two homogeneous regions and two possibly heterogeneous regions. In addition, as seen in the four-region state, applying CCA-wWAKM results in satisfying the homogeneity conditions in all the regions. Also, while CCA-WAKM1, CCA-WAKM1,2, and CCA-WAKM1,2,3 provide four homogeneous regions, CCA-WAKM1,3 identifies a possibly heterogeneous region along with the three homogeneous regions.

As seen in Figure 5, using WAKM to identify five regions results in identifying two possibly heterogeneous regions,
while the other three regions satisfy the homogeneity conditions. Among CCA-WAKM implementation options, CCA-WAKM1, CCA-WAKM1,2, and CCA-WAKM1,3, provide four homogeneous regions and one possibly heterogeneous region, whereas the option CCA-WAKM1,2,3 identifies five homogeneous regions. In this case, the use of CCA-wWAKM yields identifying five homogenous regions.

In Table 7, a summary of the performances of the examined methods of regionalization in identifying the homogeneous regions for RFFA is presented. The ratio of the number of homogeneous regions identified by each regionalization option to the total number of regions identified by that option is calculated separately based on each of the measures $H_1$, $H_2$, and $H_3$. The percentage of homogeneous regions provided by each method, which are identified by all the three heterogeneity measures as homogeneous, to all the regions identified by that option can be seen in the last column of Table 7. The percentage of homogeneous regions is defined as Equation (4):

$$p_H = \frac{n_H}{n_t} \times 100$$

where $p_H$ is the percentage of homogeneous regions, $n_H$ represents the number of homogeneous regions and $n_t$ denotes the total number of regions in two, three, four, and five-region states ($n_t = 2 + 3 + 4 + 5 = 14$).

The results indicate that among the methods and their implementation options, the best performance in providing
the homogeneous regions is related to CCA-wWAKM. All 14 regions identified by CCA-wWAKM in two, three, four, and five-region states are identified as homogeneous according to all the three heterogeneity measures. CCA-wWAKM shows perfect efficiency (100%) in identifying homogeneous regions in the study area.

A probable reason for the superiority of CCA-wWAKM over the other options is to use information related to all the watershed features. While in the implementation options of CCA-WAKM the multiples of watershed features are used in a linear combination to provide values of a regionalization feature, in CCA-wWAKM each watershed feature is included in the regionalization feature vector separately. In addition, the weight of each feature is determined only based on the absolute magnitude of its coefficient in the linear combination of the canonical variable $V_1$.

**Table 7** The percentage of homogeneous regions (%) identified in regionalization for two, three, four, and five regions according to the heterogeneity measure $H$

| Regionalization method | $H_1$ | $H_2$ | $H_3$ | $H_1, H_2, H_3$ |
|------------------------|-------|-------|-------|-----------------|
| WAKM                   | 92.9  | 78.6  | 78.6  | 64.3            |
| CCA-wWAKM              | 100   | 100   | 100   | 100             |
| CCA-WAKM$_1$           | 92.9  | 85.7  | 100   | 85.7            |
| CCA-WAKM$_{1,2}$       | 85.7  | 92.9  | 100   | 78.6            |
| CCA-WAKM$_{1,3}$       | 100   | 78.6  | 78.6  | 78.6            |
| CCA-WAKM$_{1,2,3}$     | 92.9  | 100   | 100   | 92.9            |

**Figure 5** The values of the heterogeneity measures $H_1$, $H_2$, and $H_3$ for five regions identified by WAKM, CCA-wWAKM, and CCA-WAKM.
After CCA-wWAKM, CCA-WAKM\(_{1,2,3}\) displays the best performance by identifying 13 homogeneous regions based on all three heterogeneity measures \(H\). In fact, according to Figures 2–5, all the regions identified by this option are homogeneous according to \(H_2\) and \(H_3\), and only in the three-region state, a region is identified as possibly heterogeneous by \(H_1\). As a result, the efficiency of this option in identifying homogeneous regions in the study area can be estimated at 93%. For the option CCA-WAKM\(_1\), the efficiency is equal to 86% approximately. Also, for the options CCA-WAKM\(_{1,2}\) and CCA-WAKM\(_{1,3}\), among 14 identified regions, 11 regions are shown to be homogeneous on the basis of all three measures \(H\). Thus, the efficiency of these options in identifying the homogeneous regions is approximately 79%. According to Hosking & Wallis (1997), the heterogeneity measure \(H_1\) is more sensitive to heterogeneity of regions in comparison with the measures \(H_2\) and \(H_3\). However, as seen in Table 6, this is not observed for CCA-WAKM\(_{1,3}\), because in the regionalization states with four and five regions, in each regionalization, CCA-WAKM\(_{1,3}\) identified one region including a group of sites with a relatively high value of standard deviation of the statistics L-skewness and L-kurtosis. L-skewness and L-kurtosis play key roles in definitions of the measures \(H_2\) and \(H_3\) (Hosking & Wallis 1997). By applying WAKM, 9 homogeneous regions were identified among 14 regions, which yields the lowest efficiency among the methods used for regionalization in this study (about 64%). This means that all options of the proposed method are more efficient in providing homogeneous regions than WAKM. Moreover, the CCA-wWAKM is superior to all four CCA-WAKM options by providing 100% efficiency.

Figure 6 | Geographical dispersion of the sites over two regions.
The better performance of CCA-WAKM\textsubscript{1,2,3} in comparison with other options of CCA-WAKM is because of adding the second and third canonical variables $V_2$ and $V_3$ to regionalization feature vectors. $V_2$ and $V_3$ show higher correlations with L-moment ratios L-skewness and L-kurtosis which play important roles in calculation of the heterogeneity measures $H_2$ and $H_3$. According to Table 7, by adding $V_2$ and $V_3$ to the regionalization feature vectors, homogeneity of the regions is improved to some extent based on the heterogeneity measures $H_2$ and $H_3$. Of course, as seen in Table 6, the effect of $V_2$ is more considerable than $V_3$ on the homogeneity improvement.

In general, the results of calculating the heterogeneity indices $H$ for the regions show that CCA-wWAKM and all the four CCA-WAKM options outperform WAKM in effectively identifying homogeneous regions for the study area. Therefore, all of the implementation options of the proposed method can be used as effective alternatives to common regionalization methods in order to improve the homogeneity of the identified regions. Among the CCA-WAKM implementation options, the CCA-WAKM\textsubscript{1,2,3} outperforms the other options in terms of the percentage of homogeneous regions. In addition, CCA-wWAKM is the optimum options, because of its excellent performance in identifying the regions satisfying homogeneity condition completely. This method even outperforms CCA-WAKM\textsubscript{1,2,3}. Of course, the difference between the results of these two options is only related to the measure $H_1$ in the second region that exceeds the threshold $H = 1$ in the three-region state for CCA-WAKM\textsubscript{1,2,3}.

After the homogeneity assessment, the size of regions, i.e., the total number of flood data contained by each region, was evaluated. Also, the assignment of the sites

![Figure 7](http://iwaponline.com/hr/article-pdf/50/4/1076/584685/nh0501076.pdf)

**Figure 7** Geographical dispersion of the sites over three regions.
to the regions was studied. For this evaluation, CCA-WAKM$_{1,2,3}$ is selected as the best option for representing CCA-WAKM, due to its better performance in identifying homogeneous regions than other CCA-WAKM options. For both WAKM and CCA-wWAKM, it is not needed to choose an optimal option, because there is only one implementation option for each of them.

Figures 6–9 show the assignment of the sites to the regions identified by WAKM, CCA-WAKM$_{1,2,3}$, and CCA-wWAKM, respectively. According to the figures, the geographical contiguity in the regions identified by the CCA-wWAKM is more considerable than that in the regions provided by the single WAKM and CCA-WAKM$_{1,2,3}$. In other words, delineating the crisp geographical boundaries for the regions identified by CCA-wWAKM in the study area is more feasible compared to the regions provided by the other options. In fact, for this study area and the selected features, implementation of regionalization using the CCA-wWAKM method results in the assignment of greater weights to the features related to the geographical location of the sites, especially the longitude. It should be noted that all the positive and negative coefficients of the watershed features in the first canonical variable of the watershed features used as clustering feature weights are squared in the Euclidean distance. Thus, only the absolute magnitude of these coefficients or weights affects the clustering.

In addition, according to the number of sites assigned to the regions, the dispersion of the sites across the regions provided by the single WAKM and CCA-wWAKM is more balanced than that across the regions identified by CCA-WAKM$_{1,2,3}$. Figure 10 displays the sizes of the regions identified by the regionalization methods in terms of the number

![Figure 8](http://iwaponline.com/hr/article-pdf/50/4/1076/584685/nh0501076.pdf)

Figure 8 | Geographical dispersion of the sites over four regions.
of flood data recorded in each region (station-years) for two, three, four, and five regions.

As seen in Figure 10, the number of regions with a size lower than 100 station-years identified by CCA-WAKM_{1,2,3} is greater than those identified by the other methods. Considering the fact that the average flood data record length in the selected sites is about 23 years, a region with size greater than 100 station-years can include at least four watersheds with the average flood data record length. It should be noted that in RFFA, the main goal is to increase the reliability of flood estimates by increasing the number of flood data pooled from several sites in the homogeneous regions. The regionalization that yields the identification of small regions may not be so useful to achieve this goal and so, it cannot be the optimal option for RFFA. Indeed, RFFA is characterized by a trade-off between the size of the region (i.e., number of flood data in station-years) and its homogeneity: usually, the higher the size of the identified pooling group of sites, the higher the expected heterogeneity. Moreover, the target-size of the region depends on the return period associated with the target flood quantile (see, e.g., Cunnane 1988; Jakob et al. 1999). Therefore, the use of the CCA-wWAKM method in three, four, and five-region states seems to provide better results than the CCA-WAKM_{1,2,3}.

Considering the excellent performance of CCA-wWAKM in the identification of homogeneous regions with appropriate spatial proximity and more balanced assignment of the sites to the regions, it seems that this method can be selected as the optimal option for regionalization of watersheds in Sefidrud basin for RFFA.

Figure 9 | Geographical dispersion of the sites over five regions.
CONCLUSIONS

In this study, a hybrid regionalization method was proposed by combining CCA and WAKM clustering algorithms in order to increase the homogeneity of the identified regions for RFFA. Performances of the methods in the Sefidrud basin in northern Iran were evaluated based on ASW as a cluster validity and the measures $H_1$, $H_2$, and $H_3$ as the heterogeneity measures.

According to the values of the ASW cluster validity index, the quality of clustering performed by all the options of the proposed method was higher than that of the clustering done by WAKM.

Also, the homogeneity assessment of the regions based on the values of the heterogeneity measures indicated that CCA-wWAKM and all four implementation options of CCA-WAKM were more efficient in identifying homogeneous regions than WAKM. Among the CCA-WAKM options, CCA-WAKM1,2,3 with an efficiency of 93% in the identification of homogeneous regions showed the best performance and so, it was identified as the optimal CCA-WAKM option. However, the best performance among all the options discussed was related to CCA-wWAKM. All the identified regions by CCA-WWAKM in two- to five-region states satisfied the homogeneity conditions completely. Thus, this option resulted in a 100% efficiency in providing the homogeneous regions. Therefore, CCA-wWAKM can be regarded as the optimal option for identifying the most homogeneous regions, among all the options discussed in this study.

The evaluation of the assignment of the sites to the regions identified by the regionalization methods showed that the geographical proximity of the sites in the regions identified by the CCA-wWAKM is clearer than those of the other options and methods. This may be because of the high weight of the geographical features, especially the longitude in comparison with the other features, in regionalization by CCA-wWAKM.

In addition, it was observed that the distribution of the sites across the regions identified by CCA-wWAKM is more balanced in terms of the number of flood data contained by the regions compared to that of CCA-WAKM. In

![Figure 10](image-url) Sizes of the identified regions (station-years) by WAKM, CCA-wWAKM, and the selected CCA-WAKM option.
In fact, the use of CCA-WAKM\textsuperscript{1,2,3} in some cases led to the identification of large regions (in terms of station-years) along with small regions. Identifying small regions is not so desirable for RFFA because it is not possible to provide reliable flood estimates for the sites of these regions. Thus, while both CCA-WAKM and CCA-wWAKM seem efficient in identifying homogeneous regions in comparison with WAKM, CCA-wWAKM can be the more appropriate option for regionalization of watersheds in Sefidrud basin.

As the final remark, it should be noted that examining the effectiveness of the proposed method in case studies with the larger total area makes it possible to apply the regionalization methods for a higher number of regions. Of course, it depends on the target-size of the region, which is related to the return period considered for flood quantile estimation. Also, access to a higher number of watershed features can lead to a more accurate judgment about the advantages and disadvantages of the proposed method.

REFERENCES

Acreman, M. C. & Sinclair, C. D. 1986 Classification of drainage basins according to their physical characteristics; an application for flood frequency analysis in Scotland. Journal of Hydrology 84 (34), 365–380.

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

Ahani, A., Mousavi Nadoushani, S. S. & Moridi, A. 2018 A feature weighting and selection method for improving the homogeneity of regions in regionalization of watersheds. Hydrological Processes 32 (13), 2084–2095.

Asong, Z. E., Khalig, M. N. & Wheeler, H. S. 2015 Regionalization of precipitation characteristics in the Canadian prairie provinces using large-scale atmospheric covaraites and geophysical attributes. Stochastic Environmental Research and Risk Assessment 29 (3), 875–892.

Basu, B. & Srinivas, V. V. 2014 Regional flood frequency analysis using kernel-based fuzzy clustering approach. Water Resources Research 50 (4), 3295–3316.

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.

Bhaskar, N. R. & O’Connor, C. A. 1989 Comparison of method of residuals and cluster analysis for flood regionalization. Journal of Water Resources Planning and Management 115, 793–808.

Burn, D. H. 1989 Cluster analysis as applied to regional flood frequency. Journal of Water Resources Planning and Management 115 (5), 567–582.

Burn, D. H. 1990 An appraisal of the ‘region of influence’ approach to flood frequency analysis. Hydrological Sciences Journal 35 (2), 149–165.

Burn, D. H. 1997 Catchment similarity for regional flood frequency analysis using seasonality measures. Journal of Hydrology 202, 212–230.

Burn, D. & Goel, N. K. 2000 The formation of groups for regional flood frequency analysis. Hydrological Sciences Journal 45 (1), 97–112.

Burn, D. H., Zrinji, Z. & Kowalchuk, M. 1997 Regionalization of catchments for regional flood frequency analysis. Journal of Hydrologic Engineering 2 (2), 76–82.

Castellarin, A., Burn, D. H. & Brath, A. 2001 Assessing the effectiveness of hydrological similarity measures for flood frequency analysis. Journal of Hydrology 241, 270–285.

Cavadias, G. S. 1990 The canonical correlation approach to regional flood estimation. In: Regionalization in Hydrology (Proceedings of the Ljubljana Symposium). IAHS, Wallingford, pp. 171–178.

Cavadias, G. S., Ouarda, T. B. M. J., Bobée, B. & Girard, C. 2001 A canonical correlation approach to the determination of homogeneous regions for regional flood estimation of ungaged basins. Hydrological Sciences Journal 46 (4), 499–512.

Chen, L.-H. & Hong, Y.-T. 2012 Regional Taiwan rainfall frequency analysis using principal component analysis, self-organizing maps and L-moments. Hydrology Research 43 (3), 275–285.

Cunnane, C. 1988 Methods and merits of regional flood frequency analysis. Journal of Hydrology 100, 289–290.

Dalrymple, T. 1960 Flood Frequency Analysis. Water Supply Paper 1543A, US Geological Survey, Washington, DC, USA.

Di Prinzio, M., Castellarin, A. & Toth, E. 2011 Data-driven catchment classification: application to the pub problem. Hydrology and Earth System Sciences 15 (6), 1921–1935.

Farsadnia, F., Rostami Kamrood, M., Moghaddam Nia, A., 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.

GREGHYS 1996a Presentation and review of some methods for regional flood frequency analysis. Journal of Hydrology 186, 63–84.

GREGHYS 1996b Inter-comparison of regional flood frequency procedures for Canadian rivers. Journal of Hydrology 186, 85–103.

Hall, M. J. & Minns, A. W. 1999 The classification of hydrologically homogeneous regions. Hydrological Sciences Journal 44 (5), 693–704.

Hartigan, J. A. & Wong, M. A. 1979 Algorithm as 136: a K-means clustering algorithm. Journal of the Royal Statistical Society 28 (1), 100–108.

Hosking, J. R. M. 1990 L-moments: analysis and estimation of distributions using linear combinations of order statistics. Journal of the Royal Statistical Society Series B 52, 105–124.
Hosking, J. R. M. & Wallis, J. R. 1993 Some statistics useful in regional frequency analysis. *Water Resources Research* **29** (2), 271–281.

Hosking, J. R. M. & Wallis, J. R. 1997 *Regional Frequency Analysis – An Approach Based on L-Moments*. Cambridge University Press, New York, USA.

Hotelling, H. 1936 Relations between two sets of variates. *Biometrika* **28** (3/4), 321–377.

Ilorme, F. & Griffis, V. W. 2013 A novel procedure for delineation of hydrologically homogeneous regions and the classification of ungaged sites for design flood estimation. *Journal of Hydrology* **492**, 151–162.

Jakob, D., Reed, D. W. & Robson, A. J. 1999 Choosing a pooling-group. In: Flood Estimation Handbook, vol. 3, *Statistical Procedures for Flood Frequency Estimation*. Institute of Hydrology, Wallingford, UK, pp. 153–180.

Jin, Y., Liu, J., Lin, L., Wang, A. & Chen, X. 2017 Exploring hydrologically similar catchments in terms of the physical characteristics of upstream regions. *Hydrology Research* 49 (5), 1467–1483.

Jingyi, Z. & Hall, M. J. 2004 Regional flood frequency analysis for the Gan-Ming River basin in China. *Journal of Hydrology* **296** (1–4), 98–117.

Lin, G.-F. & Chen, L.-H. 2006 Identification of homogeneous regions for regional frequency analysis using the self-organizing map. *Journal of Hydrology* **324** (1–4), 1–9.

Mosley, M. P. 1981 *Delimitation of New Zealand hydrological regions*. *Journal of Hydrology* **49**, 173–192.

Nathan, R. J. & McMahon, T. A. 1990 Identification of homogeneous regions for the purposes of regionalisation. *Journal of Hydrology* **121**, 217–238.

Nezhad, M. K., Chokmani, K., Ouarda, T. B. M. J., Barbet, M. & Bruneau, P. 2010 Regional flood frequency analysis using residual kriging in physiographical space. *Hydrological Processes* **24** (15), 2045–2055.

Ouarda, T. B. M. J., Girard, C., Cavadias, G. S. & Bobée, B. 2001 Regional flood frequency estimation with canonical correlation analysis. *Journal of Hydrology* **254** (14), 157–173.

Ouarda, T. B. M. J., Bâ, K. M., Diaz-Delgado, C., Carsteaun, A., Chokmani, K., Gingras, H., Quentin, E., Trujillo, E. & Bobée, B. 2008 Intercomparison of regional flood frequency estimation methods at ungaged sites for a Mexican case study. *Journal of Hydrology* **348** (1–2), 40–58.

Ramachandra Rao, A. & Srinivas, V. V. 2006a Regionalization of watersheds by hybrid-cluster analysis. *Journal of Hydrology* **318** (1–4), 37–56.

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

Ramachandra Rao, A. & Srinivas, V. V. 2008 *Regionalization of Watersheds – An Approach Based on Cluster Analysis*, Vol. 58 (Water Science and Technology Library). Springer, The Netherlands.

Razavi, T. & Coulibaly, P. 2013 Classification of Ontario watersheds based on physical attributes and streamflow series. *Journal of Hydrology* **493**, 81–94.

Reed, D. W., Jakob, D., Robson, A. J., Faulkner, D. S. & Stewart, E. J. 1999 Regional frequency analysis: a new vocabulary. In: *Hydrological Extremes: Understanding, Predicting, Mitigating* (Proceedings of IUGG 99 Symposium Birmingham, July 19). IAHS, Wallingford, pp. 237–243.

Ribeiro-Correa, J., Cavadias, G. S., Clément, B. & Rouselle, J. 1995 Identification of hydrological neighborhoods using canonical correlation analysis. *Journal of Hydrology* **173** (14), 71–89.

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.

Sadri, S. & Burn, D. H. 2011 A fuzzy C-means approach for regionalization using a bivariate homogeneity and discordancy approach. *Journal of Hydrology* **401**, 231–239.

Srinivas, V. V., Tripathi, S., Ramachandra Rao, A. & Govindaraju, R. S. 2008 Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering. *Journal of Hydrology* **348** (1–2), 148–166.

Srinivasava Raju, K. & Nagesh Kumar, D. 2008 Classification of Indian meteorological stations using cluster and fuzzy cluster analysis, and Kohonen artificial neural networks. *Hydrology Research* **38** (3), 303–314.

Tasker, G. D. 1982 Comparing methods of hydrologic regionalization. *Water Resources Bulletin* **18**, 965–970.

Toth, E. 2013 Catchment classification based on characterisation of streamflow and precipitation time series. *Hydrology and Earth System Sciences* **17** (5), 1149–1159.

Viglione, A., Laio, F. & Claps, P. 2007 A comparison of homogeneity tests for regional frequency analysis. *Water Resources Research* **43** (3), W03428.

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

Wiltshire, S. E. 1986 Regional flood frequency analysis II: multivariate classification of drainage basins in Britain. *Hydrological Sciences Journal* **31**, 335–346.

Xie, P., Lei, X., Zhang, Y., Wang, M., Han, I. & Chen, Q. 2018 Cluster analysis of drought variation and its mutation characteristics in Xinjiang province, during 1961–2015. *Hydrology Research* **49** (4), 1016–1027.

Zrinji, Z. & Burn, D. H. 1994 Flood frequency analysis for ungaged sites using a region of influence approach. *Journal of Hydrology* **153** (1–4), 1–21.

First received 13 January 2019; accepted in revised form 7 May 2019. Available online 3 June 2019.