A multiscale time-space approach to analyze and categorize the precipitation fluctuation based on the wavelet transform and information theory concept

Kiyoumars Roushangar, Vahid Nourani and Farhad Alizadeh

ABSTRACT

The present study proposed a time-space framework using discrete wavelet transform-based multiscale entropy (DWE) approach to analyze and spatially categorize the precipitation variation in Iran. To this end, historical monthly precipitation time series during 1960–2010 from 31 rain gauges were used in this study. First, wavelet-based de-noising approach was applied to diminish the effect of noise in precipitation time series which may affect the entropy values. Next, Daubechies (db) mother wavelets (db5–db10) were used to decompose the precipitation time series. Subsequently, entropy concept was applied to the sub-series to measure the uncertainty and disorderliness at multiple scales. According to the pattern of entropy across scales, each cluster was assigned an entropy signature that provided an estimation of the entropy pattern of precipitation in each cluster. Spatial categorization of rain gauges was performed using DWE values as input data to k-means and self-organizing map (SOM) clustering techniques. According to evaluation criteria, it was proved that k-means with clustering number equal to 5 with *Silhouette coefficient* = *0.33*, *Davis–Bouldin* = *1.18* and *Dunn index* = *1.52* performed better in determining homogenous areas. Finally, investigating spatial structure of precipitation variation revealed that the DWE had a decreasing and increasing relationship with longitude and latitude, respectively, in Iran.

Key words | discrete wavelet transform (DWT), entropy concept, Iran, k-means clustering, precipitation regionalization, self-organizing map (SOM)

Kiyoumars Roushangar (corresponding author)
Vahid Nourani
Farhad Alizadeh
Department of Water Resources Engineering,
Faculty of Civil Engineering,
University of Tabriz,
29 Bahman Ave., Tabriz,
Iran
E-mail: *kroshangar@yahoo.com*

INTRODUCTION

Assessment of precipitation variation over a large area (e.g., Iran) could provide valuable information for water resources management and engineering issues, particularly in a changing climate. The impact of global warming on different water cycle components is strongly variable across the globe and causes increases in average global precipitation, evaporation, and runoff (Clark *et al.* 1999; Pechlivanidis *et al.* 2017; Salvia *et al.* 2017; Sattari *et al.* 2017; Wei *et al.* 2017; Ba *et al.* 2018). Alteration of the hydrologic cycle will have significant impacts on the rate, timing, and distribution of rain, evaporation, temperature, snowfall, and runoff, the main causes of change in the accessibility of water resources (Mishra *et al.* 2009). The example of precipitation variation in Iran (during 1966–2005) could be referred to the rate of the significant decreasing trends in annual precipitation that varied from (−)1.999 mm/year in the northwest to (+)4.261 mm/year in the west of Iran. The significant negative trends mainly occurred in the northwest of Iran. These negative trends can affect agriculture and water supply of the regions. On the contrary, no significant trends were detected in the eastern, southern, and central parts of the country (Tabari & Hosseinzadeh Talaee 2011; Raziei 2017). By considering the high spatial and temporal variability of precipitation and frequent dry

doi: 10.2166/nh.2018.143

periods, the increasing water demands for growing population as well as for industry and economic development, including irrigation, aggravating water scarcity makes it difficult for a rationale water management. Hence, determination of sub-regions according to different precipitation regimes is important for water resources management and land use planning.

In recent decades, some studies have focused on studying precipitation across Iran (Domroes *et al.* 1998; Dinpashoh *et al.* 2004; Modarres 2006; Soltani *et al.* 2007; Raziei *et al.* 2008; Modarres & Sarhadi 2009; Tabari & Hosseinzadeh Talaee 2011). Domroes *et al.* (1998) applied principal component analysis (PCA) and cluster analysis (CA) on mean monthly precipitation of 71 stations and classified the precipitation regimes into five different sub-regions. On the other hand, applying the PCA and CA to 12 variables selected from 57 candidate variables for 77 stations distributed across the entire country, Dinpashoh *et al.* (2004) divided the country into seven climate sub-regions. Rainfall climates in Iran were also analyzed by Soltani *et al.* (2007) using monthly precipitation time series from 28 main sites. To determine regional climates, a hierarchical CA was applied to the autocorrelation coefficients at different lags, and three main climatic groups were found. Tabari & Hosseinzadeh Talaee (2011) analyzed trend over different sub-regions of Iran during 1966–2005. Raziei *et al.* (2008) analyzed the spatial distribution of the seasonal and annual precipitation in western Iran using data from 140 stations covering the period 1965–2000. Applying the precipitation concentration index (PCI), the intra-annual precipitation variability was also studied. The results suggest that five homogenous sub-regions can be identified based on different precipitation regimes. Modarres & Sarhadi (2009) performed spatial and temporal trend analysis of the annual and 24-hr maximum rainfall of a set of 145 precipitation gauging stations of Iran during the period of 1955–2000. The study showed that the annual rainfall is decreasing at 67% of the stations while the 24-hr maximum rainfall is increasing at 50% of the stations.

Wavelet analysis (WA), which has been widely applied in hydrology and hydrogeology, is capable of elucidating the localized characteristics of non-stationary time series both in temporal and frequency domains (Nourani *et al.* 2009, 2015; Kisi & Shiri 2012; Danandeh Mehr *et al.* 2015; Karimi *et al.* 2016; Danesh-Yazdi *et al.* 2017), and it is just suitable for hydrologic time series analyses. The wavelet entropy, combined by WA and information theory, is an important concept of describing the variability and complexity of hydrologic time series with non-stationary and multi-temporal characteristics (Zunino *et al.* 2007). It is used to first analyze a time series by WA, such as continuous wavelet transform (CWT) and multi-resolution analysis, and then calculate the entropy measures, mainly including Shannon entropy (Jaynes 1957), mutual information (Molini *et al.* 2006), and relative entropy (Abramov *et al.* 2005). Various studies have manifested the better performance of wavelet entropy in analyzing the variability and complexity of hydrologic variables compared with traditional methods (Simpson's index, McIntosh index, Berger–Parker index, Brillouin index, etc.) (Mishra *et al.* 2009; Brunsell 2010).

The proposed technique combines discrete wavelet transform (DWT) based multiscale entropy approach with k-means and self-organizing map (SOM) clustering techniques. The discrete wavelet multiscale entropy (DWE) which is a measure of the degree of order/disorder of the signal and carries information associated with multi-frequency signal, can provide useful information about the underlying dynamic processes associated with the signal and can help in precipitation-based studies (Cazelles *et al.* 2008). Therefore, this study tried to develop a precipitation-based regionalization-based DWE approach. In this study, the DWE method was applied to monthly precipitation data observed at 31 rain gauges in Iran. Higher entropy reflects more random and complicated systems and vice versa. Traditional entropy measures usually provide inaccurate or incomplete descriptions of climatic systems which generally operate over multi-resolution scales (Li & Zhang 2008). DWT was used to decompose each of the observed precipitation time series using the Daubechies (db) wavelet to capture the multiscale variability of the precipitation based on wavelet coefficients. Next, these wavelet coefficients for each scale are used to obtain the entropy for the respective scales (Sang 2012; Agarwal *et al.* 2016). The spatial organization of this multiscale variability in terms of DWE is identified using clustering methods.

MATERIAL AND METHODS

Case study and climatological dataset

This study used monthly climate data of 31 precipitation gauges all over Iran for studying precipitation regionalization (1960–2010) (Figure 1 and Table 1). Due to the variety of information involved in hydrologic processes and need to have accurate models, monthly precipitation time series was used which include various multivariate properties such as seasonality of process. Iran is a large country (approximately 1,600,000 km^2), in which climate is mostly affected by the wide latitudinal extent. Iran is located in Southwest Asia (25° to 40°N and 44° to 63°E).

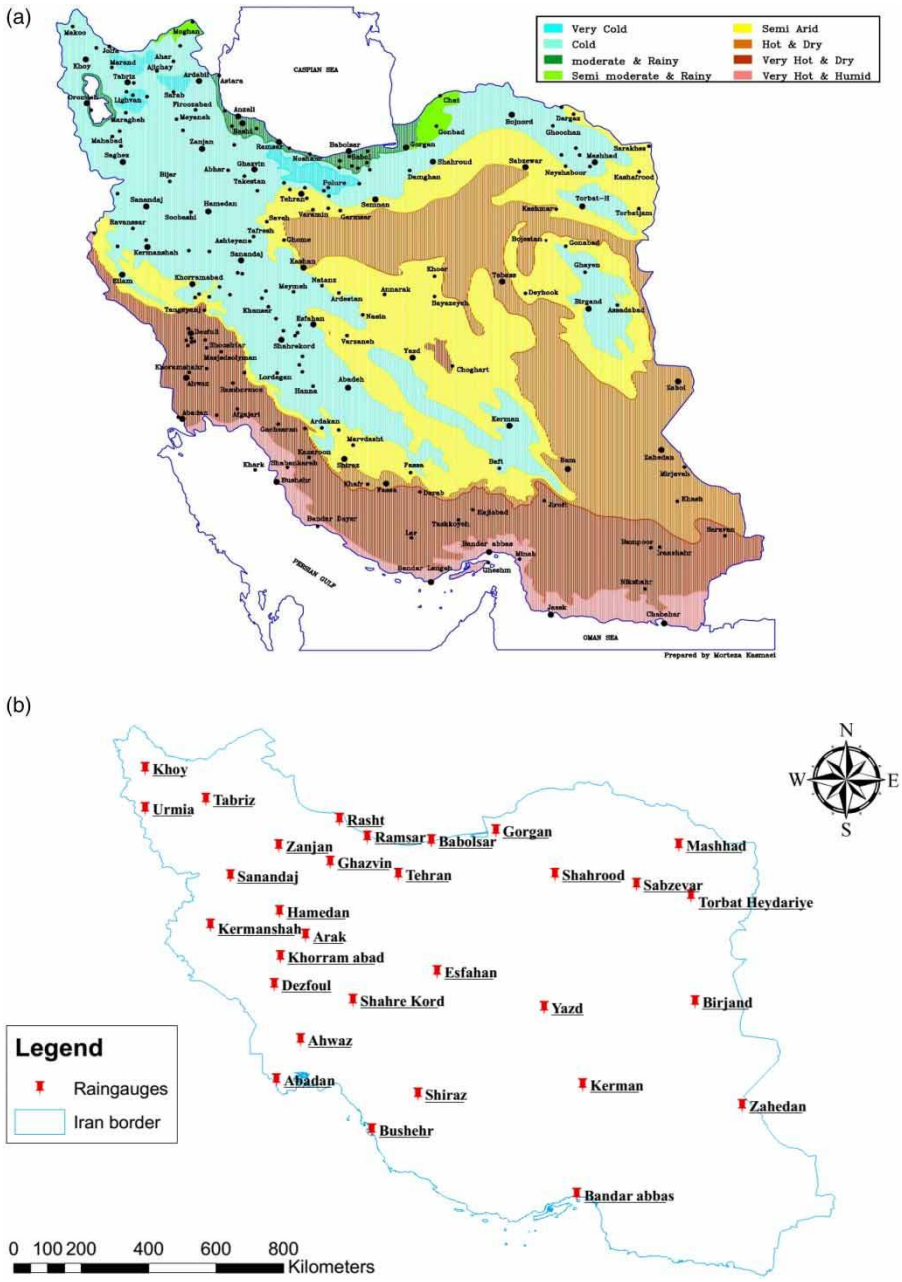

Figure 1 | (a) Climate map of Iran and (b) geographic location of rain gauges used in this study.

Table 1 | Selection of optimum number of clusters based on Dunn, Davies–Bouldin, and Silhouette indices

		Cluster number								
Clustering technique	**Validity indices**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
SOM	Silhouette index	0.19	0.28	0.27	0.28	0.34	0.30	0.21	0.18	0.09
	Davies–Bouldin index	1.59	1.10	1.45	1.40	1.20	1.48	1.7	1.62	1.35
	Dunn index	0.96	1.12	1.07	1.36	1.49	1.24	1.14	0.99	0.97
K-means	Silhouette index	0.22	0.23	0.26	0.33	0.21	0.25	0.25	0.22	0.22
	Davies–Bouldin index	1.34	1.33	1.33	1.18	1.18	1.51	1.14	1.42	1.34
	Dunn index	1.37	1.19	1.32	1.52	1.2	0.83	0.90	0.78	1.07

There are three seas in Iran, in the north the Caspian Sea and in the south the Persian Gulf and Oman Sea (Araghi *et al.* 2014). The moisture coming from the Persian Gulf is usually trapped by the Zagros Mountains. The plateau is open to the cold (dry) continental currents flowing from the northeast and the mitigating influence of the Caspian Sea is limited to the northern regions of the Alborz Mountains. The Zagros chain, which stretches from northwest to southeast, is the source of several large rivers such as the Karkheh, Dez, and Karoon. Lowland areas receive surface water from these basins and are of great importance for agricultural applications (Raziei *et al.* 2008). Iran's climate is generally recognized as arid or semi-arid with an annual average precipitation of about 250 mm; however, its climate is very diverse, with annual precipitation and temperature variation over the country (Figure 1(a)). For instance, in different areas of the country annual precipitation changes from 0 to 2,000 mm (Domroes *et al.* 1998; Dinpashoh *et al.* 2004). The Caspian Sea coastal areas along with the northern and northwestern regions of the country are subjected to higher precipitation. On the other hand, the lowest values of annual precipitation are found in the southern, eastern, and the central desert regions (Ashraf *et al.* 2013). Generally, Iran is categorized as hyper-arid (35.5%), arid (29.2%), semi-arid (20.1%), Mediterranean (5%), and wet climate (10%). Also, temperature in Iran varies widely (−20 to +50 °C) (Saboohi *et al.* 2012). On the northern edge of the country (the Caspian coastal plain) temperatures rarely fall below freezing and the area remains humid for all of the year. Summer temperatures rarely exceed 29 °C (Nagarajan 2010; Weather & Climate Information 2015). To the west, settlements in the Zagros basin experience lower temperatures, severe winters with below zero average daily temperatures and heavy snowfall. The eastern and central basins are arid and have occasional deserts. Average summer temperatures rarely exceed 38 °C (Nagarajan 2010). The coastal plains of the Persian Gulf and Gulf of Oman in southern Iran have mild winters, and very humid and hot summers (Figure 1(a)). The dataset applied in this study was provided by the Iran Meteorological Organization (http://www.irimo.ir).

Discrete wavelet transform (DWT)

The wavelet transform (WT) is a popular method and a very precise method for time series processing (Kisi & Shiri 2011; Nourani *et al.* 2014; Farajzadeh & Alizadeh 2017). While the general theory behind WT is quite analogous to that of the short-time Fourier transform (STFT), WT allows for a completely flexible window function (called the mother wavelet), which can be changed over time based on the shape and compactness of the signal. Given this property, WT can be used to analyze the time-frequency characteristics of any kind of time series. In recent years, WT has been widely used for the analysis of many hydro-meteorological time series (Adamowski *et al.* 2009; Partal 2010; Shiri & Kisi 2010; Nourani *et al.* 2013, 2015; Mehr *et al.* 2014). As the mother wavelet moves across the time series during the WT process, it generates several coefficients that represent the similarity between the time series and the mother wavelet (at any specific scale). There are two main types of WT: continuous and discrete. Use of the CWT can generate a large number of (often unnecessary) coefficients, making its use and interpretation more complicated. On the other hand, the DWT method simplifies the transformation process while still providing a very effective and precise analysis,

since DWT is normally based on the dyadic calculation. DWT coefficients can be calculated by the following equation (Partal 2010):

$$W_\psi(a, b) = \frac{1}{(2)^{a/2}} \sum_{t=0}^{N-1} x(t)\psi\left(\frac{t}{2^a} - b\right) \quad (1)$$

where 2^a represents the dyadic scale of the DWT. Applying DWT to a time series decomposes that time series into two ancillary time series shape components, called the approximation (A) and detail (D) components. Component A comprises the large-scale, low-frequency component of the time series, while component D represents the small-scale, high-frequency component.

Signal de-noising with wavelets

De-noising a signal using WT is based on the observation that in many signals (e.g., rainfall signals) energy is mostly concentrated in a small number of wavelet dimensions. The coefficients of these dimensions are relatively large compared to the other dimensions or to noise, which has its energy spread over a large number of coefficients. Hence, by setting to zero, the coefficients smaller than a certain threshold, noise can nearly be optimally eliminated while preserving the important information of the original signal (Donoho 1995). Because amplitude de-noising is performed instead of frequency de-noising, the low frequency noise can also be suppressed (Nourani & Partoviyan 2017).

To de-noise a signal using WT, the detail coefficients are thresholded, since they represent mainly noise. One way to threshold the detail coefficients is to use 'soft' thresholding. In this case, the thresholded details are given by the following equation:

$$D_{th}(i) = \begin{cases} sign(D(i))(|D(i)| - \lambda) & \text{if} |D(i)| > \lambda \\ 0 & \text{if} |D(i)| \leq \lambda \end{cases} \quad (2)$$

In Equation (2), λ and $D(i)$ $(j = 1, 2, \ldots, M)$ indicate threshold value and absolute value of detailed sub-series at ith resolution level, respectively. The algorithm to de-noise a signal $f(k)$ corrupted by a noise signal $n(k)$ can be summarized by the following three steps:

1. Apply DWT to a noisy signal to obtain approximations $A(i)$ and details $D(i)$.
2. Apply a thresholding technique to detail coefficients $D(i)$ to obtain the thresholded coefficients $D_{th}(i)$.
3. Transform the signal back based on $A(i)$ and $D_{th}(i)$ to obtain the de-noised signal (reconstruction).

According to Donoho (1995), in the case of white Gaussian noise, the threshold (λ) can be estimated as follows (Donoho 1995):

$$\lambda = \sigma\sqrt{2\log(N)} \quad (3)$$

where N is the length of the signal and σ is the noise level, which is calculated as $\sigma = \text{MAD}/0.6745$; and MAD is the median absolute value of the details coefficients estimated for the first level.

Time series decomposition via the discrete wavelet transform

The conventional discrete WA of time series was performed on each rain gauge using the multilevel 1-D wavelet decomposition function in MATLAB (MATLAB Wavelet Toolbox). This produces the WT of the time series of the interest at all dyadic scales. The monthly precipitation input time series are all one-dimensional. Decomposing the time series using specified filters (wavelet and scaling functions) produces two types of coefficients: the approximation or residual, and detail vectors (Chou 2007). These coefficients resulted from the convolution of the original time series with a low-pass filter and a high-pass filter. The low-pass filter is the scaling function and the high-pass filter is the wavelet function. The convolutions of time series with the low-pass filter produced the approximation coefficients, which represent the large-scale or low-frequency components of the original time series. Convolutions with the high-pass filter produced the detail coefficients, which represent the low-scale or high-frequency components (Bruce *et al.* 2002). The process of time series decomposition was repeated multiple times, decomposing the original time series into several different lower-resolution components (Partal 2010). The detail and approximation coefficients produced from the time series

decomposition were then reconstructed since they are merely intermediate coefficients. These have to be re-adjusted to the entire one-dimensional signal in order to enable the investigation of their contribution to the original time series (Dong *et al.* 2008). This contribution may be reflected in the different time scales such as intra-annual, inter-annual, decadal, and multi-decadal.

Selection of an appropriate wavelet function poses significant challenges and is governed largely by the problem at hand and some of the distinctive properties of the wavelet function such as (i) its region of support and (ii) the number of vanishing moments (Maheswaran & Khosa 2012).

The region of support implies the length span of the given wavelet, which in turn affects its feature localization capabilities as it is understandable that a long and widely distributed wavelet function will calculate the instantaneous process amplitude while, at the same time, spanning a wider window of the underlying process resulting in a high degree of averaging of the process states. Vanishing moment, on the other hand, limits the wavelet's ability to suitably represent polynomial behavior or information in a time series. For example, the db2 wavelet encodes polynomials with two coefficients, i.e., a process having one constant and one linear time series component, and the db3 wavelet encodes a process having a constant, linear, and quadratic time series components. Within each family of wavelets are wavelet subclasses distinguished by their respective number of coefficients and the number of vanishing moments, as discussed below (Maheswaran & Khosa 2012).

The db wavelets were used in this study because they are commonly used mother wavelets for the DWT in hydro-meteorological wavelet-based studies (Mehr *et al.* 2013). The db wavelets provide compact support with extreme phase and highest number of vanishing moments for a given support width (Vonesch *et al.* 2007), indicating that the wavelets have non-zero basis functions over a finite interval, as well as full scaling and translational orthonormality properties (Popivanov & Miller 2002; de Artigas *et al.* 2006). These features are very important for localizing events in the time-dependent signals (Popivanov & Miller 2002). These properties are unique and cannot be found in other mother wavelets (i.e., Haar, Coife, Symlet, etc.). For the period of 612 months (51 years), in order to avoid unnecessary levels of time series decomposition in these larger datasets, the number of decomposition levels had to be determined first. This number is based upon the number of data points, as well as the mother wavelet used. The highest decomposition level should correspond to the data point at which the last subsampling becomes smaller than the filter length (de Artigas *et al.* 2006). There are several recommended methods to determine the most appropriate number of decomposition levels, of which one of the most commonly used is given by the following equation (de Artigas *et al.* 2006; Araghi *et al.* 2014):

$$L = \frac{\log\left(\frac{n}{2v-1}\right)}{\log(2)} \tag{4}$$

where L is the number of decomposition levels, n is the number of data points in the time series and v is the number of vanishing moments of the db mother wavelet. In MATLAB, v is equal to the type number of the db. Smoother db wavelets (db5–db10) were then tried for each monthly time series. Smoother wavelets are preferred here because the trends are supposed to be gradual and represent slowly changing processes. Smoother wavelets should be better at detecting long-term time-varying behavior (good frequency-localization properties) (Adamowski *et al.* 2009). In addition to this, several trend studies used smoother db mother wavelets (e.g., Kallache *et al.* (2005) used least asymmetric LA(8); de Artigas *et al.* (2006) used db7). The border conditions were also taken into consideration when performing the DWT. This is because for time series with a limited length, convolution processes cannot proceed at both ends of the time series since there is no information available outside these boundaries (Su *et al.* 2011). This is referred to as the border effect (Su *et al.* 2011). As a result, an extension at both edges is needed. Border extensions that are commonly used are zero-padding, periodic extension, and symmetrization – all of which have their drawbacks, due to the discontinuities introduced at both ends of the time series (de Artigas *et al.* 2006; Su *et al.* 2011). The default extension method used in MATLAB is symmetrization, which assumes that time series outside the original support can be recovered by symmetric boundary replication (de Artigas *et al.* 2006). Zero-padding pads the time series with zeros beyond the original support of the wavelet; periodic padding assumes that time series can

be recovered outside of the original support by periodic extension (de Artigas *et al.* 2006). The inverse discrete wavelet transform (IDWT) was then computed to ensure perfect signal reconstruction. There were three main parameters to determine for the DWT used in this study: (i) the appropriate type of db wavelet; (ii) the best method for time series border extension; and (iii) the most appropriate number of decomposition levels. In order to determine the smooth mother wavelet, optimal level of decomposition level, and the extension mode to be used in the time series analysis for each data type and dataset, two criteria were used. The first criterion used was proposed by de Artigas *et al.* (2006): all three extension modes for each db wavelet were employed in order to determine the extension method, and the db type, that would produce the lowest mean relative error (*MRE*). The MRE was calculated using Equation (5) (Popivanov & Miller 2002; de Artigas *et al.* 2006):

$$MRE = \frac{1}{n}\sum_{1}^{n}\frac{|a_j - x_j|}{|x_j|} \tag{5}$$

where x_j is the original time series value of a time series whose number of records is n, and a_j is the approximation value of x_j. The second criterion used in this study is based on the relative error (e_r). Each of the extension modes for each of the smooth db wavelets was examined in order to determine the combination (of border condition and the mother wavelet) that would produce the lowest approximation Mann–Kendall Z-value relative error (e_r). The computation of the relative error was done using the following equation:

$$e_r = \frac{|Z_a - Z_o|}{|Z_o|} \tag{6}$$

where Z_a is the MK Z-value of the last approximation for the decomposition level used, and Z_o is the MK Z-value of the original time series. For the monthly time series, the MREs of the different border conditions did not show substantial differences. The differences in the relative errors were also more noticeable among the different border extensions and the different db wavelets. Since the monthly time series have 612 months of records, and according to the optimal *MRE* and e_r values, they could be decomposed up to six levels, which correspond to 64 months.

Discrete wavelet-based multiscale entropy approach (DWE)

This study proposed an approach based on hybrid DWT, entropy and k-means models to investigate the variation and regionalize the precipitation in Iran. Figure 2 shows the schematic of modeling in this study. The monthly time series of rain gauges used in this study were firstly pre-processed using DWT. For this end, Daubechies mother wavelet (db) and proper related parameters were selected for each precipitation time series.

Furthermore, DWE was used to quantify the variability and complexity of monthly precipitation processes. In the information theories, the Shannon entropy (H) is calculated as (Brunsell 2010):

$$H(x) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) \tag{7}$$

where $p(x_i)$ is the probability density function (PDF) used to describe the random characters of variable x with the length of n. H is a measure of information; more information results in lower entropy and vice versa. Therefore, bigger H value presents more disordered and complicated precipitation processes. When using the measure of DWE, the H value is calculated based on dyadic DWT results, and Equation (8) is used to compute the PDF, which is estimated

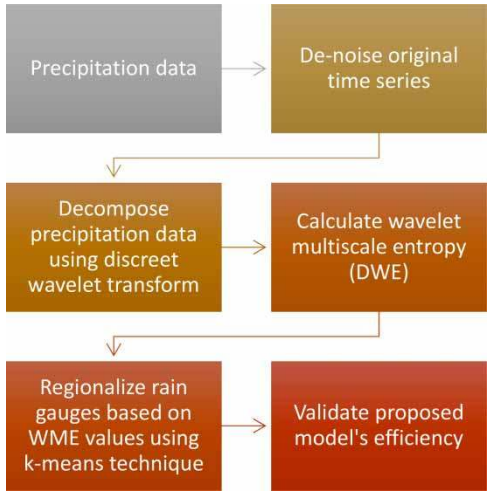

Figure 2 | Schematic of proposed DWE-based analysis and regionalization of rain gauges in this study.

according to the wavelet energy (i.e., variance) (Sang 2012):

$$P_E(m,n) = \frac{E(m,n)}{E(m)} = \frac{T(m,n)^2}{\sum_m (T(m,n))^2} \quad (8)$$

The entropy of a random variable is a measure of the uncertainty of the random variable; it is a measure of the amount of information required on average to describe the random variable (Termini & Moramarco 2016; Werstuck & Coulibaly 2016). Next, entropy values of decomposed time series (detail (D_1, $D_2 \ldots D_i$) and approximation (A_i) components) were fed into k-means approaches in order to perform spatial clustering. Different statistical evaluation criteria were used to verify the validity of clustering, which is explained below.

K-means clustering

One of the most popular clustering algorithms is the k-means method, in which the data is partitioned into k clusters, with each cluster represented by its centroid, which is the mean (weighted or otherwise) of feature vectors within the cluster (Agarwal *et al.* 2016). If N_k represents the number of feature vectors in cluster k, C_k is the mean of cluster k and X_p represents observed precipitation time series, then the centroid of each cluster is calculated using Equation (9):

$$C_k = \frac{1}{N_k} \sum_{p=1}^{N_k} X_p \quad (9)$$

The procedure follows a simple and easy way to classify a given dataset through a certain number of clusters (assume k clusters). The main idea is to define k centers, one for each cluster. The algorithm starts with the pre-defined initial number of clusters k chosen according to some criteria or some heuristic procedure. In each iteration, each cluster is assigned to its nearest cluster center according to the Euclidian distance measure between the two, and then the cluster centers (CC) are re-calculated (Rokach & Maimon 2005) until convergence of the algorithm occurs as per the defined criteria, e.g., when the algorithm exceeds the pre-defined number of iterations or when partitioning error is not going to reduce further on re-allocating cluster centroid, indicating that solution is locally optimal. The method is known for its low run time, its efficiency in clustering large datasets with numerical attributes (Rao & Srinivas 2008), and simple implementation and interpretation since no parameters (except the number of clusters) are involved. The linear complexity is also one of the reasons for the popularity of the k-means algorithm. In other words, since there are computational complexities in finding the optimal solution to the k-means clustering problem, a variety of heuristic algorithms such as Lloyd's algorithm (linear complexity) are generally used. More detailed information about the k-means clustering method can be obtained from Ball & Hall (1967) and MacQueen (1967), among others.

Self-organizing maps (SOM)

The self-organizing map is a powerful method used to explore and extract the inter-relationships of high-dimensional multivariate systems, and it is beneficial for clustering and forecasting in a widespread range of disciplines (Kohonen 1997). One of the main advantages of the SOM is its ability to extract implicit patterns from high-dimensional input dataset and classify the obtained patterns into a low-dimensional output layer, where similar inputs remain close together in the output neurons while preserving data structure (Hsu & Li 2010; Nourani *et al.* 2015). The neurons in the output layer are commonly arranged in two-dimensional grids so that the constructed topology can be visualized to give an insight into the system under investigation. The SOM has gained increasing interest and been successfully applied to hydrology and water resources management (Kalteh *et al.* 2008; Hsu & Li 2010; Nourani *et al.* 2015; Chang *et al.* 2016; Iwashita *et al.* 2018).

Evaluation criteria

In the present study, three validation metrics, namely, Davies–Bouldin index (*DBi*), Dunn index, and Silhouette coefficient (*SC*) were utilized to validate the outcome of spatial clustering via k-means technique.

In hydrology, *DBi* is a widely applied internal evaluation criterion (Davies & Bouldin 1979; Kasturi *et al.* 2003), which

is applied to distinguish the number of optimal clusters that are well-detached and well-set based on content and specification of dataset. A lower *DBi* value represents better clustering results. On the other hand, *DBi* has a disadvantage in that best information detection cannot be implied by a good reported *DBi* value.

The Dunn index's goal is to distinguish a category of clusters that are well-set, with a small variance among components of the cluster, and well detached, where the averages of the various clusters are adequately far apart when compared to the within cluster variance (Dunn 1973). A higher Dunn index's value shows better clustering outcome as it shows a well-compacted cluster (Agarwal *et al.* 2016). Computational cost increases when the number of clusters and dimensionality increases, which is a disadvantage for the Dunn index.

The SC index's goal is to show how analogous a member is to the related cluster (cohesion) in comparison to the other clusters (separation). The SC values vary from −1 to 1, where a high SC indicates that the member is well-adapted to the related cluster and insignificantly adapted to neighboring clusters.

If most members have a high SC value, then the formation of the clustering is suitable. On the other hand, if many members have a low or negative SC value, accordingly the clustering formation may have too many or too few clusters. Generally, studies have offered the applicability of SC (Hsu & Li 2010; Nourani *et al.* 2015). Nevertheless, the present study evaluated the outcome of spatial clustering based on all three indices to take advantage of them.

RESULTS AND DISCUSSION

Precipitation time series pre-processing via DWT

The precipitation time series might include a degree of noise-contamination which could influence the calculation of wavelet-based entropy values. Hence, the noise in the time series was removed by WT de-noising approach; afterwards, the DWT was applied to the de-noised time series using the chosen db mother wavelet to decompose precipitation time series into approximation and detail components.

After selecting the proper mother wavelet, boundary extensions, and decomposition level for each precipitation time series, an adequate threshold value should be chosen for the de-noising procedure. The range of threshold values within the local vicinity of the universal threshold value was acquired by Donoho's formula (Equation (3)) to determine 'appropriate threshold value' for precipitation decomposition via db mother wavelet. As an illustration, de-noised time series of rain gauge 4 (RG 4) is shown in Figure 3(a). Existence of noise in a time series can result in corruption and uncertainty by adding complexity to hydrologic time series, and the aforementioned issue becomes worse when the signal to noise ratio (SNR) value decreases. Therefore, the existence of noise in precipitation time series can significantly affect the results of the proposed model in both temporal pre-processing and spatial clustering stages. Besides, because the energy of noise mainly concentrates in small temporal scales, it has a more severe influence on the entropy values under small temporal scales than those under large temporal scales (Sang *et al.* 2011). Figure 3(b) and 3(c) show the power spectrum of original and de-noised time series, respectively. It is clearly observed that the power of de-noised time series in higher frequencies (lower temporal scales) has remarkably decreased in comparison to the original time series. Although WA analysis as a time series pre-processing method can also handle some degrees of noise included in the time series, as demonstrated in Figure 3, WD approach handle the noise in time series better, especially in higher frequencies.

Figure 4 shows the results of decomposing RG 4 precipitation time series using db9 mother wavelet and zero-padding boundary extension. Each monthly precipitation dataset was decomposed into six lower resolution levels via the DWT approach. The detail components represent the 2-month periodicity (*D1*), 4-month periodicity (*D2*), 8-month periodicity (*D3*), 16-month periodicity (*D4*), 32-month periodicity (*D5*), and 64-month periodicity (*D6*). The *A6* represents the approximation component (including the trend) at the sixth level of decomposition.

It was observed that as the transform progressed from low to high scale (short to long time scale), more boundary points became distorted due to the decimation process. In other words, as filter length increased, more points at the

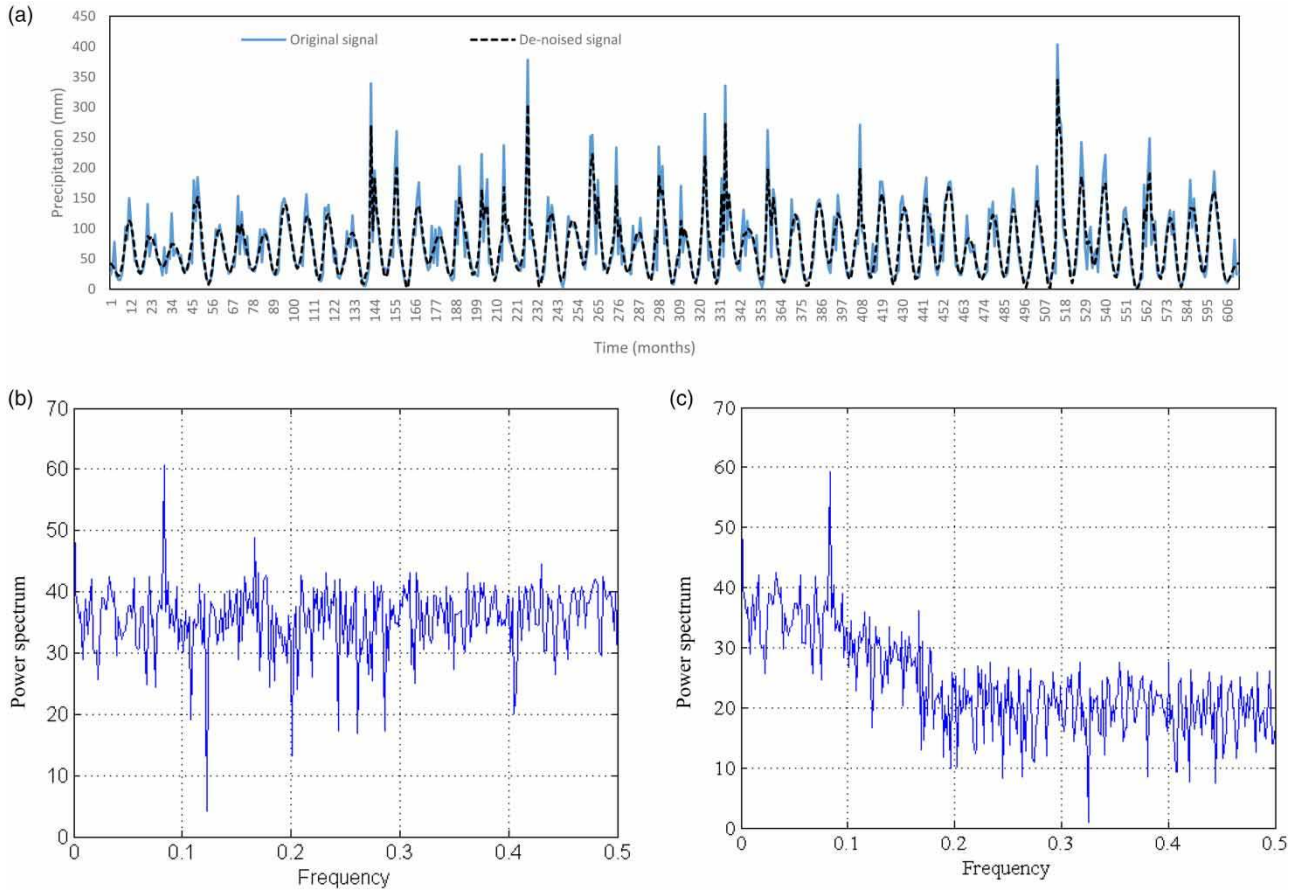

Figure 3 | (a) De-noised precipitation time series of rain gauge 4 (RG 4) via DWT, (b) power spectrum of original time series, and (c) power spectrum of de-noised time series.

boundary become affected. For higher scales (trend), the distortion becomes visibly worse. Application of boundary extension can cause inconsistency in computations of sub-series captured from DWT since it can add some uncertainties into time series (Mun 2004). This inconsistency could affect the performance of the proposed model. Hence, it was attempted to minimize the effect of applied boundary extension by using the *MRE* and e_r criteria in order to select the efficient border extension (see the 'Time series decomposition via the DWT' section).

Regionalization of rain gauges using the proposed model

At this stage, entropy-based values of the decomposed components of precipitation time series were calculated to be used as input layer of k-means. Spatial distribution of the seven entropy values of the sub-series (*A6, D6, D5 … D1*) in Iran are demonstrated in Figure 5. Based on Figure 5, highest DWE values generally were for *A6* and *D3* sub-series, whereas *D1* sub-series had the least DWE values among all values calculated. It can be observed that there are compact counter lines on the north and northern west parts of Iran. It means that DWE values of various scales change rapidly on north and northern west parts of Iran and these zones are located in rainy and cold regions of Iran. Spatial changes of DWE becomes smoother for western and southern zones which are semi-arid and arid regions of Iran. Generally, it could be stated that rapid changes of entropy pattern are observed for the northern west parts of Iran, which are mostly cold areas.

These seven values as signature of decomposed time series were used as input data to SOM and k-means in order to perform precipitation regionalization. The number

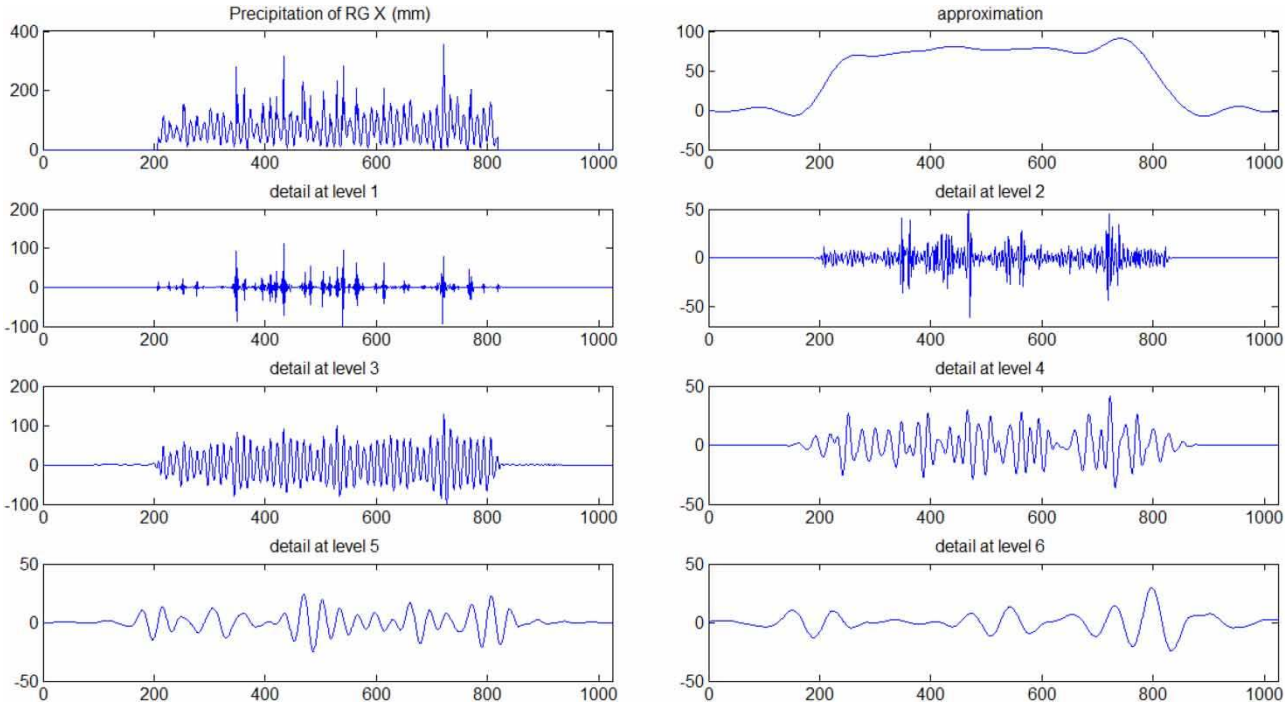

Figure 4 | Decomposition of monthly precipitation time series via db(9,6) and zero padding boundary extension for RG 4 (Ramsar).

of clusters for the dynamic features of monthly precipitation time series was determined by three validity indices' values. Table 1 shows the validity values for various numbers of DWE-based clustering approaches.

As discussed, for spatial clustering of 31 rain gauges in Iran, the DWE value of each rain gauge was used as input data of k-means clustering technique. At first, k-means approach with a 1,000 trial was trained based on DWE values. The optimal number of clusters was determined using validation indices. The clustering number 5 with $SC = 0.33$, $DBi = 1.18$, and $Dunn = 1.52$ showed a better performance in determining homogenous areas in comparison to other clustering numbers for the k-means approach. Therefore, clustering number equal to 5 was selected as the optimum value to categorize the rain gauges.

On the other hand, SOM models were used to cluster the 31 rain gauges into a visible 2-dimensional topology of regional RGL maps. For this end, map sizes of 2×2 to 10×10 were tried. The constructed topological maps coupled with related key features showed that clustering number 6 with $SC = 0.34$, $DBi = 1.20$, and $Dunn = 1.49$ led to a better performance in determining homogenous areas of precipitation variation in comparison to other clustering numbers for the SOM approach. However, there was a failure in outcome of SOM with six clusters. It was observed that two clusters had only one rain gauge, and two clusters had more than ten rain gauges. Results of k-means in means of both evaluation criteria and classification of rain gauges in various clusters proved to be better than SOM, and therefore was applied for further analysis.

Some studies took advantage of DWT-based clustering approaches as a modeling approach. For example, Hsu & Li (2010) used the WT and self-organizing map (WTSOM) framework to spatially cluster the precipitation time series. In the proposed approach, they combined the WT and a SOM neural network. WT was used to extract dynamic and multiscale features of the non-stationary precipitation time series, and SOM was employed to objectively identify spatially homogeneous clusters on the high-dimensional wavelet transformed feature space. Haar and Morlet wavelets were selected in the data pre-processing stage to preserve the desired characteristics of the precipitation data. In this study, decomposition was performed using smoother db mother wavelets (db5–db10) along with

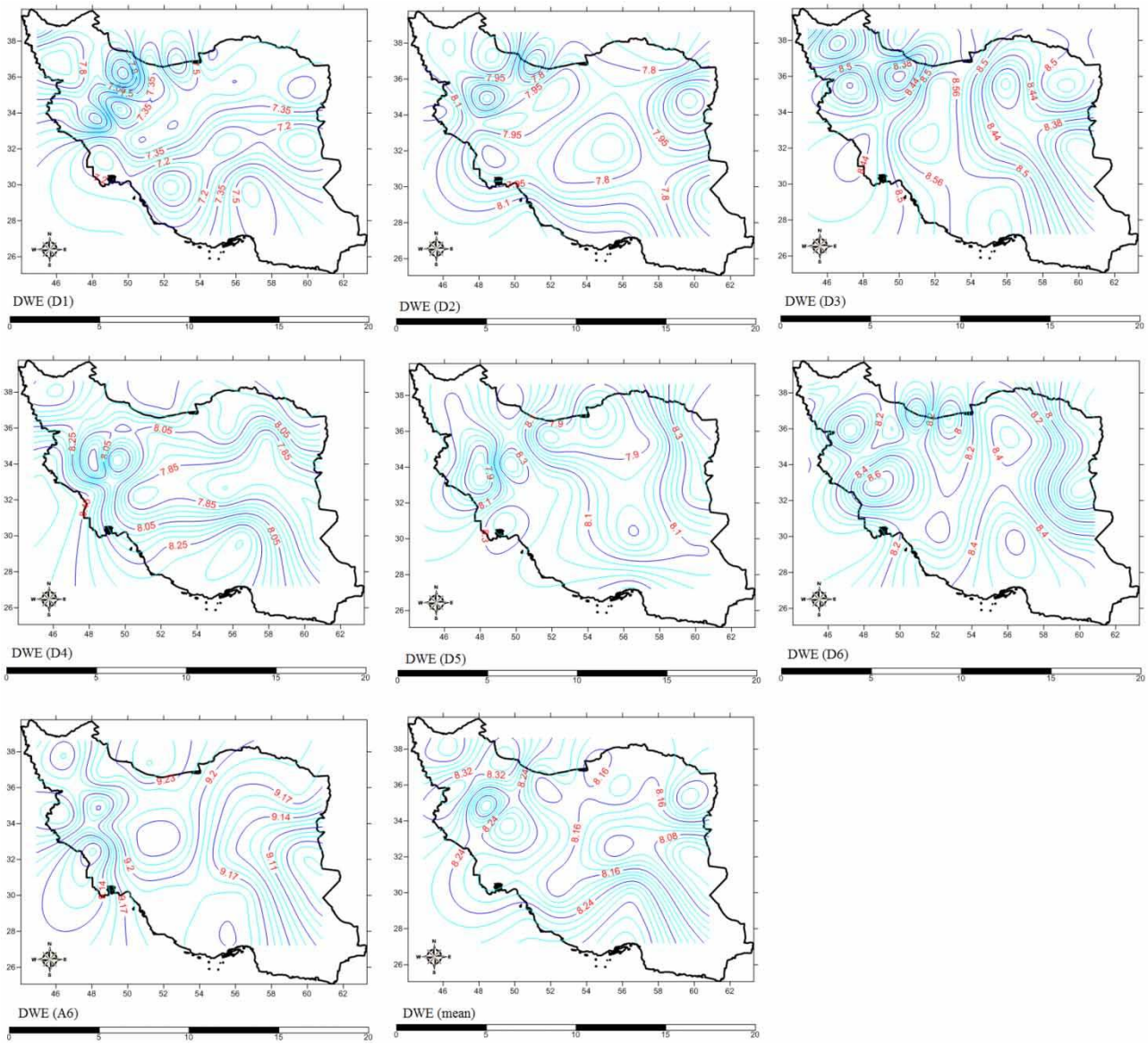

Figure 5 | Spatial distribution of discreet wavelet multiscale entropy (DWE) values over Iran.

optimum parameters. The entropy-based dynamic features of the time series could improve the performance of the clustering approach. Sub-series (i.e., Ai, Di $i = 1, 2 \ldots 6$) represent various monthly scales. Nevertheless, some of these components might not demonstrate enough correlation with rainfall original time series. For this end, DWE was calculated and used as input to k-means to perform spatial clustering.

Geographic location of rain gauges based on clustering via DWE as input into k-means approaches is demonstrated in Figure 6. Also in Figure 6, the CC based on validity indices are presented. It was seen that some of the stations in a given cluster are spread across the study area, revealing that the basis of clustering is not geographic proximity. For example, the rain gauges located near the Caspian Sea (rain gauges 19, 18, 4, and 11) with highest precipitation values and geographical proximity, are assigned to various clusters due to the differences in entropies calculated for each rain gauge. The stations in each of these clusters are further examined for any

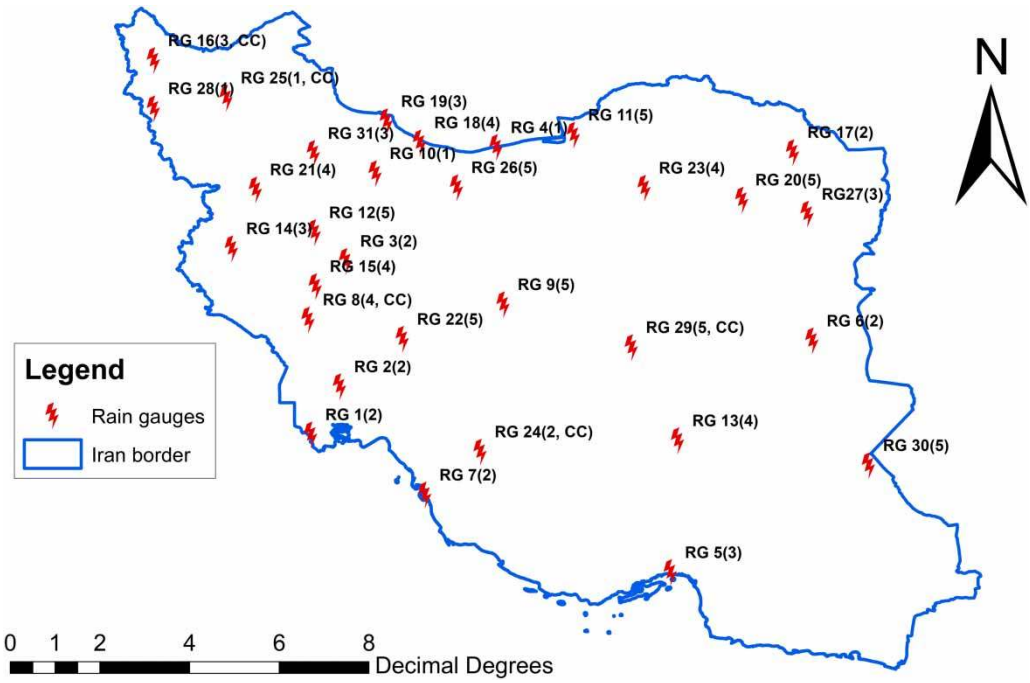

Figure 6 | Geographic location of rain gauges (RG) in each cluster along with cluster centers (CC) for WT-k-means approach.

common characteristics (in terms of multiscale entropy) they may have among themselves.

Raziei *et al.* (2008) regionalized the precipitation of the western part of Iran. They found five zones based on the behavior of precipitation. As can be seen from Figure 7, the rain gauges located in the west of Iran were placed in five different clusters. Different from results of spatial classification of rain gauges, Raziei (2017) found eight sub-regions of precipitation in Iran, namely, mountainous regime (covering Zagrom and some part of Alborz mountains), central Alborz regime, monsoonal southeastern regime, Caspian regime, northwestern regime, central-eastern regime, south and southwestern regime, and costal southeastern regime (geographic neighborhood). Also, Domroes *et al.* (1998); Modarres (2006), and Raziei (2017) separated Iran's rainfall regions into eight groups. The outcome of these studies are very analogous to each other. Similar to the present study, precipitation in the west of Iran was subjected to various precipitation changes for these studies (Figure 5). However, the results of the clustering in this study are very different, since the aforementioned studies classified the precipitation regime in Iran based on neighborhood approximates. As can be seen, the clustering shows that there is hydrologic similarity (in terms of multiscale variation of precipitation) in the clusters apart from the geographic neighborhood. It was observed that some of the rain gauges in a given cluster are spread across the study area showing that the basis of clustering is not the geographic contiguity.

As can be seen in Figure 7, the multiscale entropy values are, to a great extent, similar within any given cluster and the basis of the clustering is the entropy signature of the precipitation observed at all the rain gauges for all clusters. For example, in Cluster 1 (Figure 7(a)), the entropy signatures for all the rain gauges are similar and the peaks in the plots indicate high values of entropy, which corresponds to high variability of the precipitation features at the specific scale across time. In addition, the pattern of the entropy in a given cluster across all scales for the rain gauges is unique for that cluster (homogeneity) but also different from every other cluster.

In order to ensure a more sensible and simpler analysis, the average entropies for all clusters were used instead of entropy values of single rain gauges, and this was considered as the representative value of entropy for all clusters at a specific scale. Figure 8 shows the DWE values obtained for detail and approximation components. It was observed

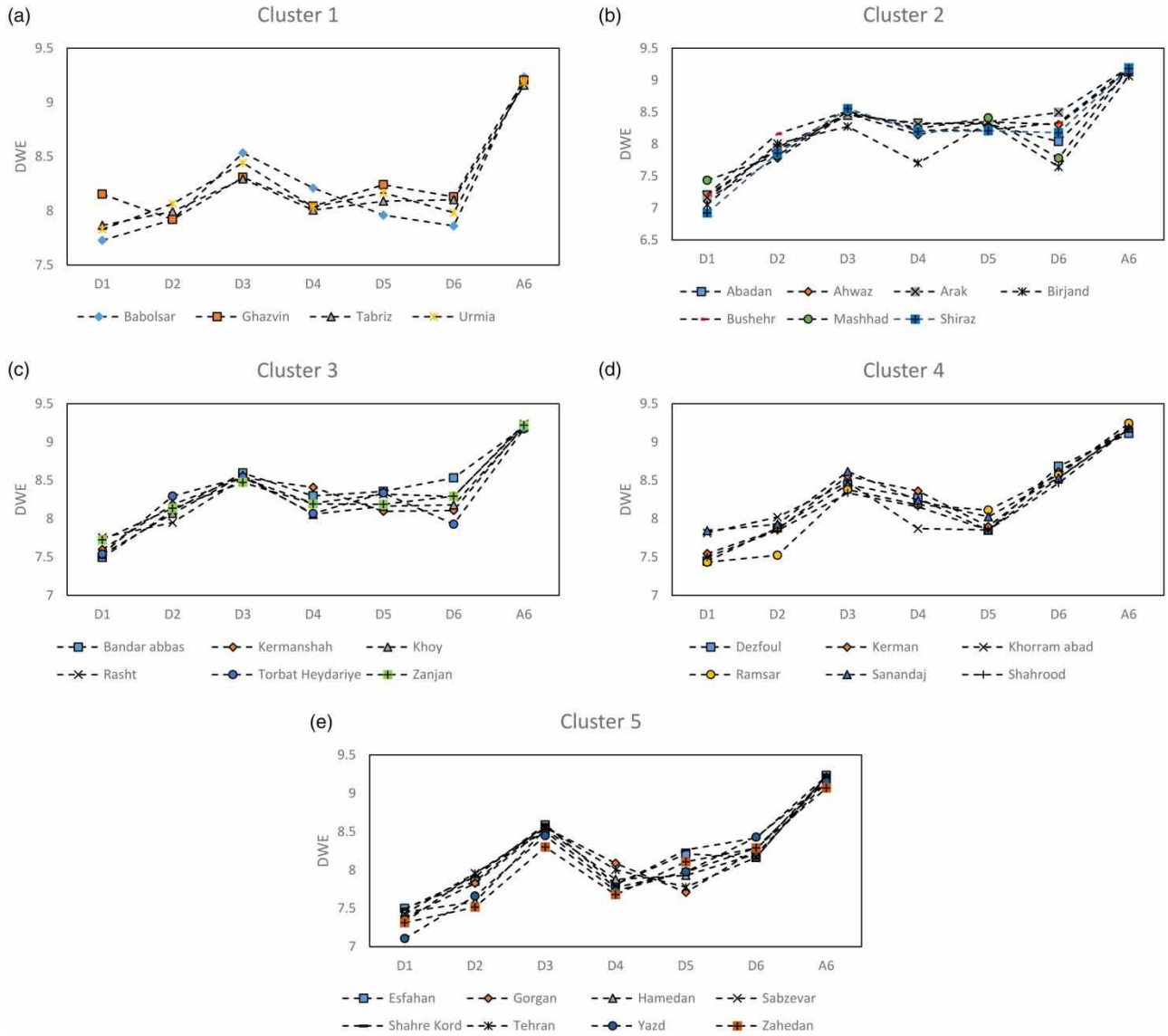

Figure 7 | Multiscale entropy values for five clusters (k-means): (a) Cluster 1, (b) Cluster 2, (c) Cluster 3, (d) Cluster 4, and (e) Cluster 5.

that entropy values of *D3* and approximation components had the highest values and lowest variation whereas *D1* had lowest entropy values with highest variation among all components. The DWE values increased from *D1* to *D3*, then decreased from *D3* to *D4*. However, variation of DWE from *D4* to *D6* and *A6* (Figures 7 and 8) was not constant. In order to prove the outcome of Figure 8, the wavelet power spectrum of central rain gauges of each cluster are presented in Figure 9. It can be observed, that for all the rain gauges, that there are very rapid changes for the period of 1 to 16 months. For the period of 1 month, mostly low powers were observed; however, for the periods up to 8 months, the change in power spectrum becomes rapid from low to high and vice versa was observed. These changes become smoother for the period of the 16-month band in comparison to the 8-month band. For the bands beyond 16 months, the changes become smoother in comparison to previous bands and also, power spectrum values are higher than the 1-month band. Therefore, it can be inferred that these entropies of *D1* to *D4* sub-series are the key variables in precipitation regionalization. Also, it can be stated that the precipitation variation is affected by

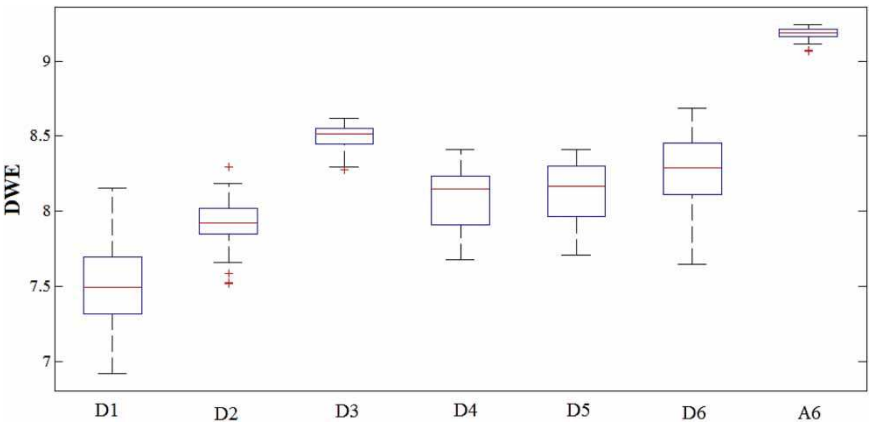

Figure 8 | Variation of DWE values for all scales of all rain gauges.

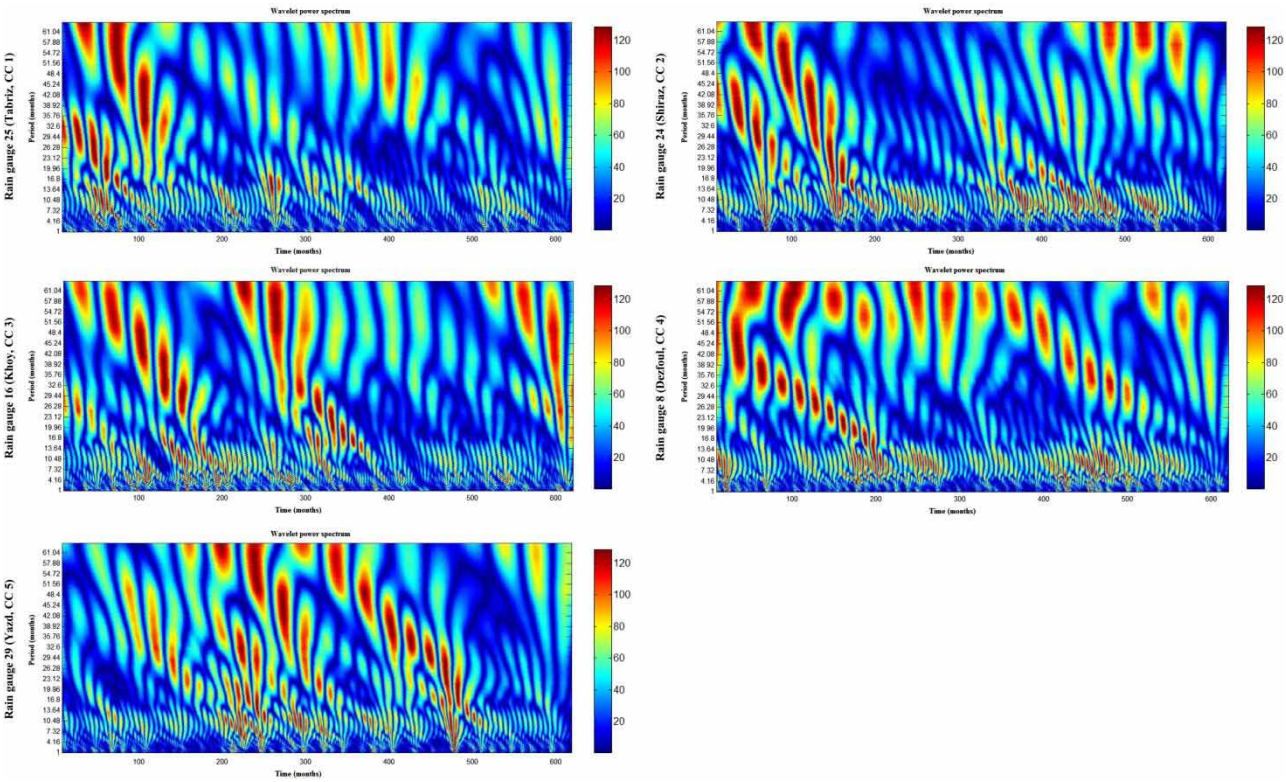

Figure 9 | Wavelet power spectrum of central rain gauges of each cluster.

different variables, such as timing, amount, and temporal distribution.

Based on these observations, three distinct bands were determined for further analysis. Band 1 considered the features up to 8-month scales (*D1*, *D2*, and *D3*). Band 2 considered the features from 8 to 16 months. The features having a scale beyond 16 months were categorized as Band 3 (*D5* and *D6*). Figure 10 shows the average normalized DWE for all the clusters at different bands (the DWE values were normalized for better comparisons). There is a clear distinction in the values of DWE for different clusters in the first two bands and, in view of this, information from the first two bands was further analyzed. The DWE of each cluster was further classified into 'High', 'Medium', and

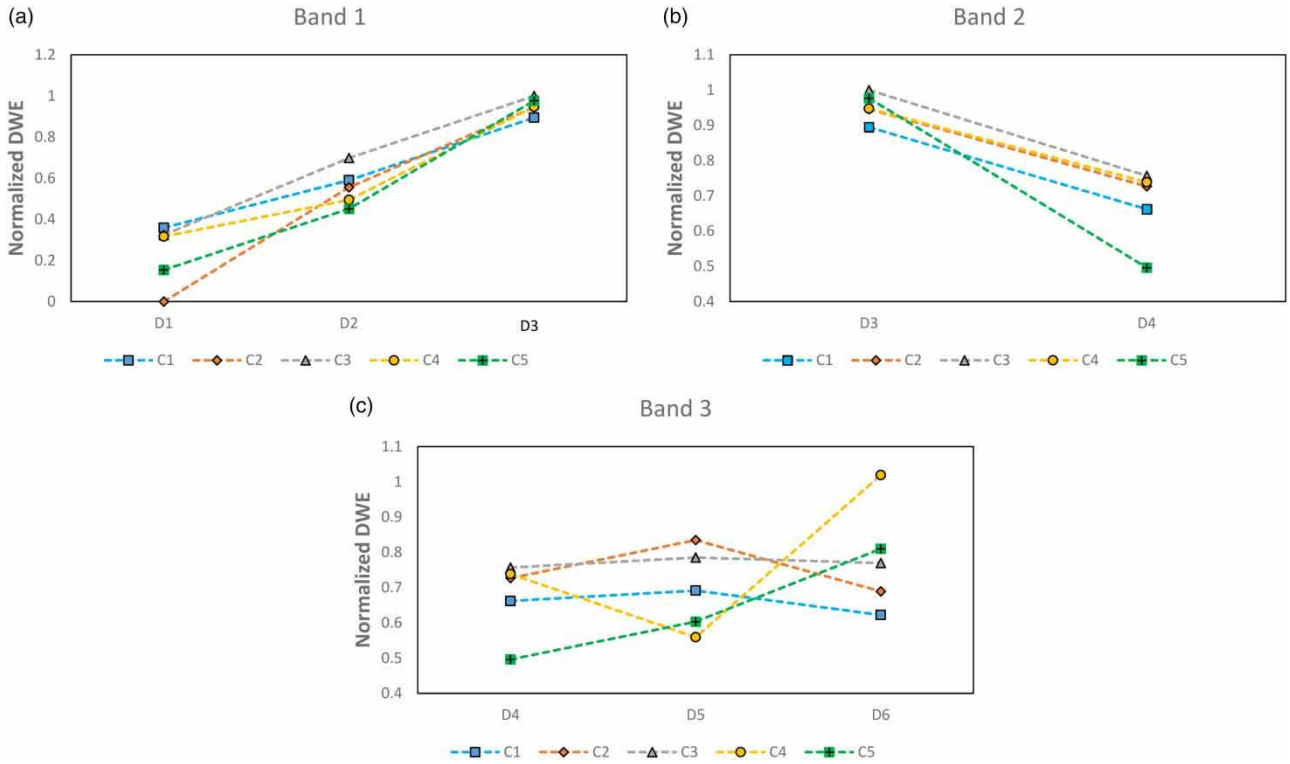

Figure 10 | Comparison of normalized DWE values for each scale for all clusters according to segregated bands: (a) Band 1: 2–8 months, (b) Band 2: 8–16 months, and (c) Band 3: 16–64 months.

'Low', by considering the condition of the individual DWE plot according to the mean level for that band. For instance, if the DWE of a cluster in a specific band fell below the mean of DWE of all clusters, then that particular cluster was assigned a signature of 'Low'. Using this classification, an entropy signature was given to each cluster based on the entropy values in the three scale-based bands. For notational simplicity, the classifications 'High', 'Medium', and 'Low' were represented by '1', '0', and '-1', respectively. This means, for example, that an entropy signature of $(0, -1)$ would indicate that the cluster had a relatively moderate entropy up to 8 months and low entropy for 8–16 months. Based on these notations, the entropy signature for each of the 14 clusters is given in Table 2.

As a further step, it was attempted to connect the DWE values at different scale-based bands to their respective mean monthly precipitation of rain gauges. Boxplots of mean monthly precipitation (Figure 11) suggest that the clusters with 'High' entropy for the scale 9–13 months (i.e., Clusters 2 and 5) had smaller precipitation values. Clusters characterized by 'Medium' or 'High' entropy for the scale 8–16 months (i.e., Clusters 1, 3, and 4) had larger precipitation values. Hence, mean monthly precipitation values and relative entropy showed an inverse relationship.

Table 2 | Entropy signature of five clusters for 31 rain gauges in Iran

Cluster no.	Comparative observation of DWE		Entropy signature
	Band 1	Band 2	
1	H	M	(1,0)
2	L	M	(−1,0)
3	H	L	(1, −1)
4	M	L	(0, −1)
5	L	L	(−1, −1)

As an important issue, the connection between the DWE with latitude and longitude was investigated to indicate the spatial structure of the precipitation variation, which is shown in Figure 12. For DWE latitude, $R^2 = 0.227$ and

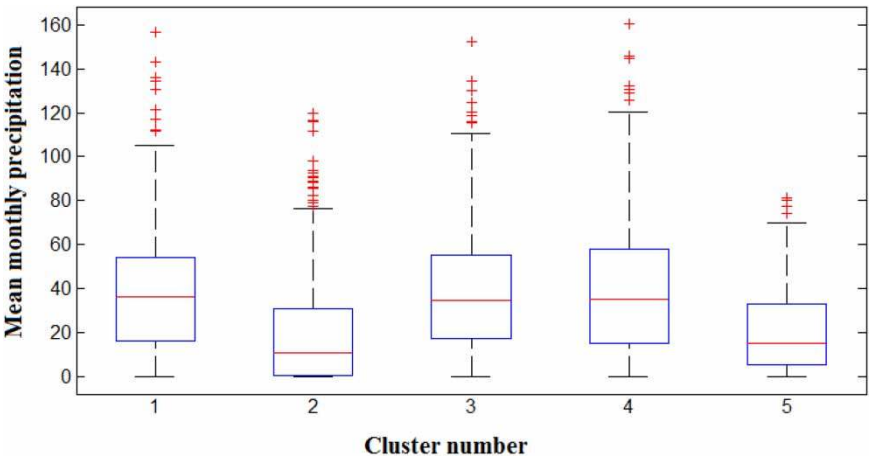

Figure 11 | Distribution of mean monthly precipitation values for the five clusters.

P-value = 0.37 (not significant) were calculated and a downward relation was observed; however, for DWE longitude, $R^2 = 0.22$ and *P*-value = 0.34 (not significant) were calculated and an upward relation was observed. For both of the relations no significant trend was detected. It can be inferred from Figure 12 that multiscale precipitation variation (DWE) possesses the latitude zonality, which implies that precipitation variability increases with the latitude from the west to the east. On the other hand, decrease of DWE values from north to south was observed.

Results showed the capability of the present methodology for precipitation regionalization. When accessibility to recorded precipitation time series is limited at the region of interest, regionalization methods might lead to incorrect results. For the case of the discrete wavelet, although the wavelet power spectrum has successfully been used for capturing hydrological time series behavior, it becomes difficult to use the wavelet spectrum in cases of limited or incomplete time series. Nevertheless, entropy provides information about the uncertainty at a given scale, which can highlight the level of variation present at that scale. Further, entropy enables the determination of least-biased probability distributions with limited time series knowledge. Entropy theory can serve as a useful approach to study hydrologic and meteorological

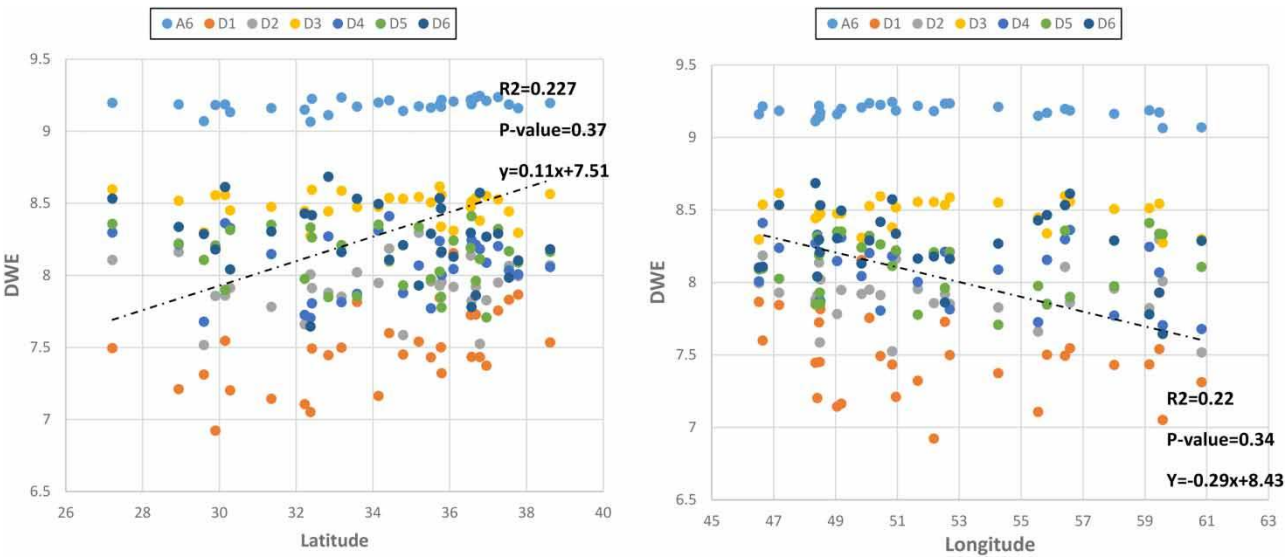

Figure 12 | Spatial structure of DWE values (in latitude and longitude directions) and related trends in Iran.

processes (Mishra *et al.* 2009; Agarwal *et al.* 2016). Sang (2012) also showed the usefulness of applying DWE in precipitation-based studies.

The obtained results are applicable in local scale, since various factors can affect the outcome of the proposed model (e.g., geographic location, precipitation variation, effect of climatic phenomena, precipitation gauges network, etc.). Due to the existence of uncertainties and various factors it is suggested to apply the proposed model for various case studies and to compare the outcome. Also, it is suggested to validate the capability of the proposed model on other hydrologic and climatic variables (i.e., evapotranspiration, temperature, runoff, etc.) with various time scales (e.g., daily, annual, etc.).

CONCLUSION

In this study, the spatio-temporal variability of monthly precipitation in Iran during 1960–2010 was investigated using DWE, and the pattern of DWE changes along with regionalization of rain gauges were further analyzed. In order to meet the objectives of this study 31 rain gauges were selected.

In order to have a correct vision of decomposing precipitation time series, smoother db mother wavelets were applied (db5–db10). Also, optimal decomposition level and boundary extension treatment were applied. In order to classify the rain gauges, SOM and k-means clustering models were used. The methodology based on the DWE approach k-means clustering technique for precipitation regionalization proved to be robust for hydrologic regionalization.

Wavelet-based multiscale entropy values showed the distinct variation of precipitation dynamics at each rain gauge and allowed for the establishment of homogeneous areas (with no prior assumptions). Most of the previous studies in precipitation regionalization delineated the rain gauges based on geographic proximity; however, the present study categorized rain gauges according to the uncertainties (entropy) in a multiscale approach. The DWE was useful circumstantial evidence in capturing the precipitation characteristics. The 31 rain gauges studied were clustered into five groups, each one having a unique DWE pattern across different time scales. Based on the pattern of mean DWE for each cluster, a characteristic signature was assigned, which provided an estimation of DWE of a cluster across scales 2–8, 8–16, and 32–64 months relative to other stations. Fluctuations in DWE at different scales in this study were related to monthly precipitation.

Results showed the capability of the present methodology for precipitation regionalization. When accessibility to recorded precipitation time series is limited at the region of interest, regionalization methods might lead to incorrect results. For the case of the discrete wavelet, although the wavelet power spectrum has successfully been used for capturing hydrological time series behavior, it becomes difficult to use the wavelet spectrum in cases of limited time series. Nevertheless, entropy provides information about the uncertainty at a given scale, which can highlight the level of variation present at that scale. Further, entropy enables the determination of least-biased probability distributions with limited time series knowledge. Entropy theory can serve as a useful approach to study hydrologic and meteorological processes (Mishra *et al.* 2009; Agarwal *et al.* 2016). Sang (2012) also showed the usefulness of applying DWE in precipitation-based studies.

REFERENCES

Abramov, R., Majda, A. & Kleeman, R. 2005 Information theory and predictability for low-frequency variability. *Journal of Atmospheric Research* **62**, 65–87.

Adamowski, K., Prokoph, A. & Adamowski, J. 2009 Development of a new method of wavelet aided trend detection and estimation. *Hydrological Processes* **23** (18), 2686–2696.

Agarwal, A., Maheswaran, R., Sehgal, V., Khosa, R., Sivakumar, B. & Bernhofer, C. 2016 Hydrologic regionalization using wavelet-based multiscale entropy method. *Journal of Hydrology* **538**, 22–32.

Araghi, A., Mousavi-Baygi, M., Adamowski, J., Malard, J., Nalley, D. & Hashemnia, S. M. 2014 Using wavelet transforms to estimate surface temperature trends and dominant periodicities in Iran based on gridded reanalysis data. *Atmospheric Research* **155**, 52–72.

Ashraf, B., Yazdani, R., Mousavi-Baygi, M. & Bannayan, M. 2013 Investigation of temporal and spatial climate variability and aridity of Iran. *Theoretical and Applied Climatology* **118** (1), 35–46.

Ba, H., Guo, S., Wang, Y., Hong, X., Zhong, Y. & Liu, Z. 2018 Improving ANN model performance in runoff forecasting by adding soil moisture input and using data preprocessing

techniques. *Hydrology Research* **49** (3), 744–760. doi: 10.2166/nh.2017.048.

Ball, G. H. & Hall, D. J. 1967 A clustering technique for summarizing multivariate data. *Systems Research and Behavioral Science* **12** (2), 153–155.

Bruce, L. M., Koger, C. H. & Jiang, L. 2002 Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction. *IEEE Transactions on Geoscience and Remote Sensing* **40** (10), 2331–2338.

Brunsell, N. A. 2010 A multiscale information theory approach to assess spatial-temporal variability of daily precipitation. *Journal of Hydrology* **385**, 165–172.

Cazelles, B., Chavez, M., Berteaux, D., Ménard, F., Vik, J. O., Jenouvrier, S. & Stenseth, N. C. 2008 Wavelet analysis of ecological time series. *Oecologia* **156** (2), 287–304.

Chang, F. J., Chang, L. C., Huang, C. W. & Kao, I. F. 2016 Prediction of monthly regional groundwater levels through hybrid soft-computing techniques. *Journal of Hydrology* **541**, 965–976.

Chou, C. M. 2007 Applying multi-resolution analysis to differential hydrological grey models with dual series. *Journal of Hydrology* **332** (1–2), 174–186.

Clark, P. U., Alley, R. B. & Pollard, D. 1999 Northern hemisphere ice-sheet influences on global climate change. *Science* **286**, 1104–1111.

Danandeh Mehr, A., Kahya, E., Şahin, A. & Nazemosadat, M. J. 2015 Successive-station monthly streamflow prediction using different artificial neural network algorithms. *International Journal of Environmental Science and Technology* **12** (7), 2191–2200.

Danesh-Yazdi, M., Tejedor, A. & Foufoula-Georgiou, E. 2017 Self-dissimilar landscapes: probing into the causes and consequences via multi-scale analysis and synthesis. *Geomorphology* **296**, 16–27.

Davies, D. L. & Bouldin, D. W. 1979 A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1** (2), 224–227.

de Artigas, M. Z., Elias, A. G. & de Campra, P. F. 2006 Discrete wavelet analysis to assess long-term trends in geomagnetic activity. *Physics and Chemistry of the Earth* **31** (1–3), 77–80.

Dinpashoh, Y., Fakheri-Fard, A., Moghaddam, M., Jahanbakhsh, S. & Mirnia, M. 2004 Selection of variables for the purpose of regionalization of Iran's precipitation climate using multivariate methods. *Journal of Hydrology* **29**, 109–123.

Domroes, M., Kaviani, M. & Schaefer, D. 1998 An analysis of regional and intra-annual precipitation variability over Iran using multivariate statistical methods. *Theoretical and Applied Climatology* **61**, 151–159.

Dong, X., Nyren, P., Patton, B., Nyren, A., Richardson, J. & Maresca, T. 2008 Wavelets for agriculture and biology: a tutorial with applications and outlook. *Bioscience* **58** (5), 445–453.

Donoho, D. H. 1995 De-noising by soft-thresholding. *IEEE Transactions on Information Theory* **41** (3), 613–617.

Dunn, J. C. 1973 A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters. *Journal of Cybernetics* **3** (3), 32–57.

Farajzadeh, J. & Alizadeh, F. 2017 A hybrid linear-nonlinear approach to predict the monthly rainfall over the Urmia Lake watershed using Wavelet-SARIMAX-LSSVM conjugated model. *Journal of Hydroinformatics* **20** (1), 246–262.

Hsu, K. C. & Li, S. T. 2010 Clustering spatial–temporal precipitation data using wavelet transform and self-organizing map neural network. *Advances in Water Resources* **33**, 190–200.

Iwashita, F., Friede, M. J., Francisco, J. & Ferreira, J. F. 2018 A self-organizing map approach to characterize hydrogeology of the fractured Serra-Geral transboundary aquifer. *Hydrology Research* **49** (3), 794–814. doi:10.2166/nh.2017.221.

Jaynes, E. T. 1957 Information theory and statistical mechanics. *Physics Review* **106**, 620–630.

Kallache, M., Rust, H. W. & Kropp, J. 2005 Trend assessment: applications for hydrology and climate research. *Nonlinear Processes in Geophysics* **12** (2), 201–210.

Kalteh, A. M., Hjorth, P. & Berndtsson, R. 2008 Review of self-organizing map in water resources: analysis, modeling, and application. *Environmental Modelling and Software* **23**, 835–845.

Karimi, S., Shiri, J., Kisi, O. & Shiri, A. A. 2016 Short-term and long-term streamflow prediction by using 'wavelet-gene expression' programming approach. *ISH Journal of Hydraulic Engineering* **22** (2), 148–162.

Kasturi, J., Acharya, J. & Ramanathan, M. 2003 An information theoretic approach for analyzing temporal patterns of gene expression. *Bioinformatics* **19** (4), 449–458.

Kisi, O. & Shiri, J. 2011 Precipitation forecasting using wavelet-genetic programming and wavelet-neuro-fuzzy conjunction models. *Water Resources Management* **25** (13), 3135–3152.

Kisi, O. & Shiri, J. 2012 Wavelet and neuro-fuzzy conjunction model for predicting water table depth fluctuations. *Hydrology Research* **43** (3), 286–300.

Kohonen, T. 1997 *Self-organizing Maps*. Springer-Verlag, Berlin.

Li, Z. W. & Zhang, Y. K. 2008 Multi-scale entropy analysis of Mississippi River flow. *Stochastic Environmental Research and Risk Assessment* **22**, 507–512.

MacQueen, J. 1967 Some methods for classification and analysis of multivariate observations. *Proceeding of Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 281–297.

Maheswaran, R. & Khosa, R. 2012 Comparative study of different wavelets for hydrologic forecasting. *Computers and Geosciences* **46**, 284–295.

Mehr, A. D., Kahya, E. & Olyaei, E. 2013 Streamflow prediction using linear genetic programming in comparison with a neuro-wavelet technique. *Journal of Hydrology* **505**, 240–249.

Mehr, A. D., Kahya, E. & Ozger, M. 2014 A gene–wavelet model for long lead time drought forecasting. *Journal of Hydrology* **517**, 691–699.

Mishra, A. K., Özger, M. & Singh, V. P. 2009 An entropy-based investigation into the variability of precipitation. *Journal of Hydrology* **370**, 139–154.

Modarres, R. 2006 Regional precipitation climates of Iran. *Journal of Hydrology: New Zealand* **45** (1), 13–27.

Modarres, R. & Sarhadi, A. 2009 Rainfall trends analysis of Iran in the last half of the twentieth century. *Journal of Geophysical Research* **114**, D03101.

Molini, A., Barbera, P. L. & Lanza, L. G. 2006 Correlation patterns and information flows in rainfall fields. *Journal of Hydrology* **322**, 89–104.

Mun, F. K. 2004 *Time Series Forecasting Using Wavelet and Support Vector Machine*. MS Thesis, National University of Singapore, Singapore.

Nagarajan, R. 2010 *Drought Assessment*. Springer Science & Business Media, New York, p. 383.

Nourani, V. & Partoviyan, A. 2017 Hybrid denoising-jittering data pre-processing approach to enhance multi-step-ahead rainfall–runoff modeling. *Stochastic Environmental Research and Risk Assessment* 1–18.

Nourani, V., Komasi, M. & Mano, A. 2009 A multivariate ANN-wavelet approach for rainfall–runoff modeling. *Water Resources Management* **23** (14), 2877–2894.

Nourani, V., Hosseini Baghanam, A., Adamowski, J. & Gebremichael, M. 2013 Using self-organizing maps and wavelet transforms for space–time pre-processing of satellite precipitation and runoff data in neural network based rainfall–runoff modeling. *Journal of Hydrology* **476**, 228–243.

Nourani, V., Hosseini Baghanam, A., Adamowski, J. & Kisi, O. 2014 Applications of hybrid wavelet – Artificial intelligence models in hydrology: a review. *Journal of Hydrology* **514**, 358–377.

Nourani, V., Alami, M. T. & Vousoughi Daneshvar, F. 2015 Wavelet-entropy data pre-processing approach for ANN-based groundwater level modeling. *Journal of Hydrology* **524**, 255–269.

Partal, T. 2010 Wavelet transform-based analysis of periodicities and trends of Sakarya basin (Turkey) streamflow data. *River Research and Applications* **26** (6), 695–711.

Pechlivanidis, I. G., McIntyre, N. & Wheater, H. S. 2017 The significance of spatial variability of rainfall on simulated runoff: an evaluation based on the Upper Lee catchment, UK. *Hydrology Research* **48** (4), 1118–1130.

Popivanov, I. & Miller, R. J. 2002 Similarity search over time-series data using wavelets. In: *Proceedings 18th International Conference on Data Engineering*, Washington, DC, pp. 212–221.

Rao, A. R. & Srinivas, V. V. 2008 *Regionalization of Watersheds: an Approach Based on Cluster Analysis*, Vol. 58. Springer Science & Business Media, New York.

Raziei, T. 2017 A precipitation regionalization and regime for Iran based on multivariate analysis. *Theoretical and Applied Climatology* **131** (3–4), 1429–1448.

Raziei, T., Bordi, I. & Pereira, L. S. 2008 A precipitation-based regionalization for Western Iran and regional drought variability. *Hydrology and Earth System Sciences* **12**, 1309–1321.

Rokach, L. & Maimon, O. (eds). 2005 Clustering methods. In: *Data Mining and Knowledge Discovery Handbook*. Springer, New York, pp. 321–352.

Saboohi, R., Soltani, S. & Khodagholi, M. 2012 Trend analysis of temperature parameters in Iran. *Theoretical and Applied Climatology* **109**, 529–547.

Salvia, K., Villarinia, G. & Vecchib, G. A. 2017 High resolution decadal precipitation predictions over the continental United States for impacts assessment. *Journal of Hydrology* **553**, 559–573.

Sang, Y. F. 2012 Wavelet entropy-based investigation into the daily precipitation variability in the Yangtze River Delta, China, with rapid urbanizations. *Theoretical and Applied Climatology* **111**, 361–370.

Sang, Y. F., Wang, D., Wu, J. C., Zhu, Q. P. & Wang, L. 2011 Wavelet-based analysis on the complexity of hydrologic series data under multi-temporal scales. *Entropy* **13**, 195–210.

Sattari, M.-T., Rezazadeh-Joudi, A. & Kusiak, A. 2017 Assessment of different methods for estimation of missing data in precipitation studies. *Hydrology Research* **48** (4), 1032–1044.

Shiri, J. & Kisi, O. 2010 Short-term and long-term streamflow forecasting using a wavelet and neuro-fuzzy conjunction model. *Journal of Hydrology* **394**, 486–493.

Soltani, S., Modarres, R. & Eslamian, S. S. 2007 The use of time series modelling for the determination of rainfall climates of Iran. *International Journal of Climatology* **27**, 819–829.

Su, H., Liu, Q. & Li, J. 2011 Alleviating border effects in wavelet transforms for nonlinear time-varying signal analysis. *Advances in Electrical and Computer Engineering* **11** (3), 55–60.

Tabari, H. & Hosseinzadeh Talaee, P. 2011 Analysis of trends in temperature data in arid and semi-arid regions of Iran. *Global and Planetary Change* **79**, 1–10.

Termini, D. & Moramarco, T. 2016 Application of entropic approach to estimate the mean flow velocity and Manning roughness coefficient in a high-curvature flume. *Hydrology Research* nh2016106. doi: 10.2166/nh.2016.106.

Vonesch, C., Blu, T. & Unser, M. 2007 Generalized Daubechies wavelet families. *IEEE Transactions on Signal Processing* **55** (9), 4415–4429.

Weather and Climate Information 2015 *Weather and Climate: Iran, Average Monthly Rainfall, Sunshine, Temperature, Humidity and Wind Speed*. World Weather and Climate Information, The Netherlands.

Wei, Q., Sun, C., Wu, G. & Pan, L. 2017 Haihe River discharge to Bohai Bay, North China: trends, climate, and human activities. *Hydrology Research* **48** (4), 1058–1070.

Werstuck, C. & Coulibaly, P. 2016 Hydrometric network design using dual entropy multi-objective optimization in the Ottawa River Basin. *Hydrology Research* nh2016344. doi: 10.2166/nh.2016.344.

Zunino, L., Perez, D. G., Garavaglia, M. & Rosso, O. A. 2007 Wavelet entropy of stochastic processes. *Physics A* **379**, 503–512.

First received 27 August 2017; accepted in revised form 16 January 2018. Available online 12 February 2018