



Regionalization of potential evapotranspiration using a modified region of influence

Marzieh Hasanzadeh Saray^{1,2} · Seyed Saeid Eslamian¹ · Björn Klöve² · Alireza Gohari¹

Received: 7 September 2019 / Accepted: 10 December 2019 / Published online: 27 December 2019
© The Author(s) 2019

Abstract

This study examined the effect of different attributes on regionalization of potential evapotranspiration (ET_p) in Urmia Lake Basin (ULB), Iran, using the region of influence (RoI) framework. Data for the period 1997–2016 from 30 weather stations were selected for the analysis. To achieve similarity between stations, climate, geographical, and statistical attributes were selected. To determine the effect of each attribute, the Shannon entropy weighting method was used. The results showed that attribute weighting had a significant impact on ET_p clustering. Among the groups studied, the most significant effect of weighting was observed in the statistical attributes category. Among all attributes, skewness coefficient (C_S) was the most useful in determining similarity between stations. Based on the results, ULB can be divided into three homogeneous regions. Proximity of weather stations did not always indicate similarity between them, but by weighting the stations in addition to weighting the attributes, more accurate estimates of ET_p in the basin were obtained. Overall, the results demonstrate potential for application of the RoI approach in regionalization of ET_p, by assigning a weight to weather stations and to influencing attributes.

1 Introduction

Water scarcity and inefficient water use are the main limiting factors for agricultural development and food production in Iran. Precise estimation of evapotranspiration (ET) is needed to increase water use efficiency. Evapotranspiration is the most critical parameter for climate and hydrological studies and one of the main components of the water balance of each region of Iran. Evapotranspiration utilizes around 60% of annual solar radiation received at the Earth's surface (Wang and Dickinson 2012; Wild et al. 2013). Apart from being involved in the energy balance, ET is a significant component of the water cycle and uses about two-thirds of the rain on Earth (Baumgartner and Reichel 1975). It also plays a crucial role in atmospheric processes, as it determines the supply of water in the atmosphere from oceans and terrestrial areas. It affects the amount and spatial distribution of global

temperature and pressure (Shukla and Mintz 1982). This can affect the incidence of heatwaves (Seneviratne et al. 2006) and rainfall processes (Zveryaev and Allan 2010), and the performance of agricultural production, especially in arid and semi-arid regions.

It is widely argued that increasing temperature as a result of climate change has a direct impact on hydrological parameters such as ET (McKenney and Rosenberg, 1993). In this regard, within the period 1960–2005, the minimum and maximum temperatures at weather stations in Urmia Lake Basin (ULB) in north-west Iran increased from 0.1 to 0.5 °C and from 0.1 to 0.3 °C, respectively. This increase in temperature is leading to a decrease in lake level by accelerating the rate of water loss from Lake Urmia, which is considered to be the leading cause of desiccation of this lake. It is also leading to an increase in evapotranspiration and crop water requirements.

In hydrological studies in arid regions, it is vital to have a good understanding of the spatial variation in ET. Analysis of homogeneous areas in terms of climate characteristics, especially ungauged areas or regions with incomplete data, can improve irrigation scheduling and result in more appropriate use of water resources. Since the ET is measured as spot values, but the temperature and ET distribution on the Earth's surface are highly variable on a spatial scale, measurement of ET offers acceptable accuracy only in small environments. It is not suitable for large environments where weather

✉ Marzieh Hasanzadeh Saray
m.hasanzadeh@ag.iut.ac.ir

¹ Department of Water Engineering, College of Agriculture, Isfahan University of Technology, Isfahan 8415683111, Iran

² Water Resources and Environmental Engineering Research Unit, Faculty of Technology, University of Oulu, Oulu, Finland

stations are less densely distributed. Given this limitation, in order to identify the spatial pattern of ET in a region, the data points need to be converted to surface data. Various types of regionalization and interpolation methods can be used to evaluate spatial changes in ET.

If weather data for many stations are studied, a statistical method for identifying homogeneous climate areas should be used. In ungauged basins, model parameters should be estimated from other sources of information. An appropriate method for setting model parameters in basins that lack data is to use model parameters from a similar hydrological basin (Merz et al. 2006). Various types of regionalization method have been introduced for transfer of parameters from a similar hydrological basin to an ungauged area. These include spatial proximity, which uses interpolation techniques based on geographical locations, or spatial distances, including Kriging, inverse distance weighting (IDW), and spline, which have been applied previously to determine spatial distributions of ET (Mardikis et al. 2005; Xu et al. 2006; Zhu et al. 2012; Kamali et al. 2015; Nam et al. 2015; Lu et al. 2016).

Clustering (hierarchical and non-hierarchical) methods have also been widely used for this purpose. Da Silva et al. (2017) used the Ward method to classify reference ET for the Amazon region, while Ramos et al. (2008) used *K*-means and the Ward algorithm for ET clustering in the Sonora River Basin in Mexico. These are examples of the geostatistics, interpolation, and clustering methods used in studies conducted worldwide and in Iran to identify homogeneous ET regions.

There are many widely used methods for estimating hydrological and climate parameters for ungauged stations, but all are usually associated with errors. The recently presented region of influence (RoI) method is the latest regionalization method for solving the problems with conventional methods and is reported to produce accurate and reliable estimates with fewer errors (Wiltshire 1986). The RoI approach was first used as an alternative method for transfer of practical information from nearby weather stations to estimate flow rate at a target station (Wiltshire 1986; Acreman and Wiltshire 1987). In this approach, each station is allowed to have a unique region that creates an area for an ungauged station, making it superior to conventional regionalization methods (Burn 1990a). Zrinji and Burn (1994) confirmed this conclusion for ungauged stations in Canada. A great variety of RoI-based applications have since been used in flood estimation (Burn 1990a, 1990b; Zrinji and Burn 1994; Tasker et al. 1996; Castellarin et al. 2001; Holmes et al. 2002; Merz and Blöschl 2005; Chiu et al. 2005; Eng et al. 2007; Tsang et al. 2011). RoI-based applications have also been used in estimation of extreme rainfall (Gaál et al. 2008a; Gaál and Kysely 2009; Bharath and Srinivas 2015; Dehghan et al. 2018a; Dehghan et al. 2018b) and in regionalization of low flow (Holmes et al. 2002). The results indicate that the RoI approach is preferable

to other methods. However, no previous study has estimated ET using this method.

The first step in regionalization is to identify and collect information that can be used to calculate the proximity and similarity between several weather stations in the desired area, which are defined as attributes. Attributes used previously in investigation using the RoI approach include predictor and geographical variables (Chiu et al. 2005), climatological and geographical characteristics (Gaál et al. 2008b), geological and physical variables (Samuel et al. 2011), and climate, geographical, and statistical attributes with their hybrids (Dehghan et al. 2018b). Merz and Blöschl (2005) and Eng et al. (2007) obtained their best estimates when they considered both predictor variables and geographical proximity. Dehghan et al. (2018b) concluded that statistical attributes, in combination with climate and geographical characteristics, gave the best estimates of quantiles in terms of low relative error, and that skewness can play a useful role in evaluation of quantiles.

The next step in regionalization is to use a tool to determine similarity between stations. In the metric space, this similarity is defined by the distance criterion. Researchers have employed different distance criteria, but the Euclidean distance metric has been used in most studies (Burn 1990a, 1990b; Holmes et al. 2002; Eslamian 2010a, 2010b; Dehghan et al. 2018a, 2018b). By applying appropriate weights for available attributes in regions without data, acceptable and reliable results can be obtained for ungauged stations (Dehghan et al. 2018a).

In previous studies on regionalization, various attributes have been found to affect the goal, depending on conditions in the target region, indicating that all attributes should not be allocated the same degree of importance. There are several techniques for determining the weight of different attributes in multiple-attribute decision-making (MADM) problems, one of which is the Shannon entropy method. Shannon (1948) introduced the concept of information entropy, defined as a measure of the degree of turbulence within a system, which can have a significant effect on the identification of practical elements and their impact. The Shannon entropy concept has been widely used in hydrology (Singh 2011).

The entropy method has been used recently for a range of purposes, including determining the significance of rain gauge stations in spatiotemporal scaling (Wei et al. 2014), field velocity distribution during flood events (Chiu and Tung 2002; Moramarco et al. 2004; Farina et al. 2014), rainfall-runoff modeling (Jowitt 1991), averaged rate of infiltration (Singh 2010a), soil moisture (Al-Hamdan and Cruise 2009; Singh 2010b), distribution of piezometric head in groundwater flow (Barbe et al. 1994), estimation of discharge (Moramarco and Singh 2001; Chiu et al. 2005), and flow and sediment concentrations (Chiu et al. 2000).

Given the strategic location of Lake Urmia in north-west Iran and the fact that it is the largest hypersaline lake in the Middle East, many studies have focused on ULB (Fazel et al. 2017; Dehghan et al. 2018a, 2018b; Haghghi et al. 2018; Akbari et al. 2019). Various studies around the world have analyzed the spatial distribution of ET, but these studies have limitations as they only explain geographical attributes, regardless of their weight in clustering. To extend the analysis, we assessed the applicability of the RoI approach according to the degree of importance and participation of each attribute in regionalization of ETp in ULB. To obtain more accurate results in regionalization, we developed a framework for appropriate weighting in regionalization of ETp. In our novel approach, weighting is based not only on geographical attributes, but also on climatological and statistical attributes. We evaluated and compared the performance of the weighted attributes using both clustering and the RoI approach.

2 Materials and methods

2.1 Study area

The analysis was based on long-term weather data for Urmia Lake Basin in north-west Iran, which lies between 35° 41'–38° 30' N and 44° 13'–47° 53' E, and is 140 km long and 40–55 km wide. The basin covers a total area of 52,000 km²,

which is approximately 3% of the total area of the entire country. Around 65% of the catchment area of Lake Urmia consists of mountainous regions, 24% of plains and foothills, and 10% is occupied by the lake itself. The basin is surrounded by the northern part of the Zagros Mountains, the southern slopes of the Sabalan Mountains, and the northern, western, and southern hills of Mount Sahand. Lake Urmia, with a maximum depth of 16 m, is classified as a shallow lake, which increases its vulnerability to evaporation. The annual evaporation rate from the lake surface is estimated to be between 0.98 and 1.2 m, reflecting the dry climate in ULB. For the present analysis, daily weather data from 30 stations (see Table 1) were obtained from the Meteorological Organization and Water Resources Management Company of Iran. The historical data covered 20 years, 1997–2016. Figure 1 shows the spatial distribution of selected stations.

2.2 Determination of potential evapotranspiration

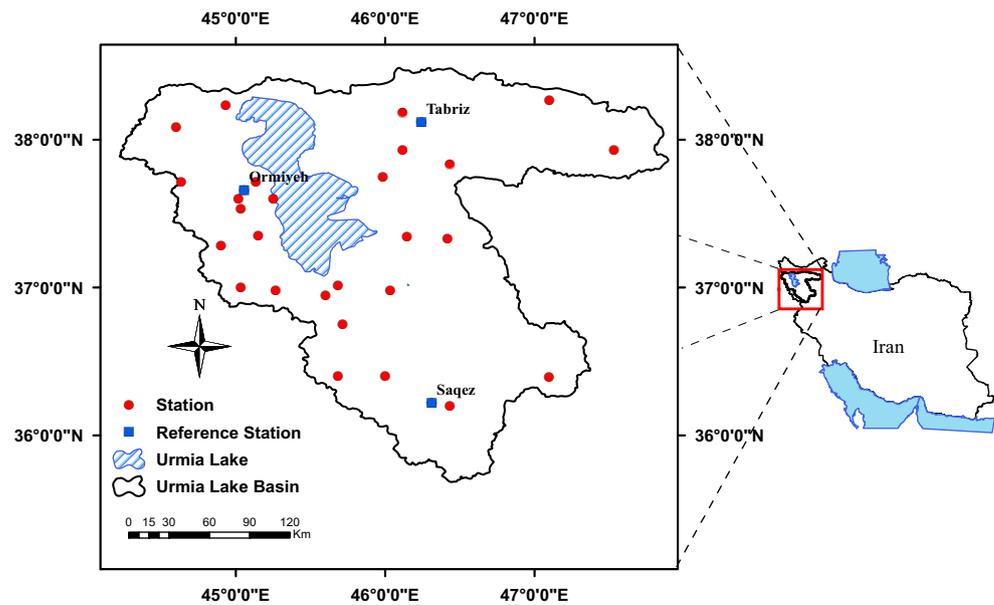
Potential evapotranspiration can be computed from meteorological data. Numerous studies around the world have found that the adapted FAO Penman-Monteith (FAO-56 PM) model (Eq. 1) is the most accurate method for estimating ETp. This method is widely used and recommended as the standard method for determining ETp from meteorological data (Allen et al. 1998).

Table 1 The geographical location of the selected stations in Urmia Lake Basin

Station number	Station name	Longitude (°)	Latitude (°)	Elevation (m)	Station number	Station name	Longitude (°)	Latitude (°)	Elevation (m)
1	Abajaloo Sofly	45.13	37.72	1290	16	Mehmandar	45.60	36.95	1350
2	Azarshahr	45.98	37.75	900	17	Pey Jik	45.27	36.98	1375
3	Barazin	47.10	38.27	1962	18	Pey Qalee	45.03	37.00	1500
4	Camp Ormiye	45.03	37.53	1381	19	Pol Adinan	46.43	36.20	1460
5	Chahriq Olya	44.60	38.08	1600	20	Qasemloo	45.15	37.35	1380
6	Dashkhane	45.68	37.02	1278	21	Sad Shahid Kazemi	46.00	36.40	1437
7	Golmankhaneh Ab Shirin	45.25	37.60	1320	22	Sahand	46.12	37.93	1641
8	Golmankhaneh Ab Shur	45.02	37.60	1320	23	Sahlan	46.12	38.18	1330
9	Hashemabad Bibakran	44.90	37.28	1570	24	Saqez*	36.22	46.31	1523
10	Kavlan	45.68	36.40	1520	25	Sarab	47.53	37.93	1682
11	Liqvan	46.43	37.83	2200	26	Tabriz*	46.24	38.12	1361
12	Maghanjiq	46.42	37.33	1500	27	Takab	47.10	36.40	1817
13	Mahabad	45.72	36.75	1352	28	Tazekand Miandoab	46.03	36.98	1290
14	Maraqe	46.15	37.35	1344	29	Urmia*	45.06	37.66	1328
15	Marz Seroo	44.63	37.72	1640	30	Yalqozaghaj	44.93	38.23	1300

*Reference stations

Fig. 1 The geographical location of Urmia Lake Basin and the selected meteorological stations



$$ET_P = \frac{0.408\Delta (R_n - G) + \gamma \left(\frac{900}{T + 273} \right) u_2 (e_s - e_a)}{\Delta + \gamma(1 + 0.34 u_2)} \quad (1)$$

where ET_P is the potential crop evapotranspiration (mm day^{-1}), Δ is the slope of the saturation vapor pressure function ($\text{kPa } (^\circ\text{C})^{-1}$), R_n is the net radiation ($\text{MJ m}^{-2} \text{day}^{-1}$), G is the soil heat flux density ($\text{MJ m}^{-2} \text{day}^{-1}$), γ is a psychrometric constant ($\text{kPa } (^\circ\text{C})^{-1}$), T is the mean temperature ($^\circ\text{C}$), u_2 is the wind speed at 2 m height (m s^{-1}), e_s is the saturation vapor pressure (kPa), e_a is the actual vapor pressure (kPa), and $e_s - e_a$ is the saturation vapor pressure deficit (kPa). The factor $0.408 = 1/\lambda$ ($\lambda =$ latent heat of vaporization in MJ kg^{-1}) converts units from $\text{MJ m}^{-2} \text{day}^{-1}$ to mm day^{-1} . In this study, all parameters necessary for computing potential evapotranspiration with the FAO-56 PM method were calculated according to the procedure developed by Allen et al. (1998).

2.3 Selection of attributes

The information used to calculate the similarity between different weather stations in the study area was divided into different attributes. The types of attributes used in the regionalization method play a key role in the success of further regionalization steps. A wide range of statistical, climatological, and geographical information was used in this study to effectively transfer data from the basin stations to reference stations. The geographical proximity of stations is considered to be a suitable indicator for similarity values of evapotranspiration. However, simple geo-proximity between the two points cannot be interpreted as similarity of stations. As the value of the attributes increases, the probability of creating dependent variables also increases. To increase the precision, climatological

and statistical attributes were also considered in this study. Among the geographical attributes, longitude (x), latitude (y), and height above sea level (h) were selected. The set of climatological site attributes consisted of average daily wind speed (WS), average daily relative humidity (RH), and average daily temperature (T). Coefficient of variation (C_V), coefficient of skewness (C_S), and the ratio of C_V to C_S were selected as the statistical attributes.

2.4 Weighting approach

Weighting coefficients represented the relative importance of the attributes for each of the selected stations. Since all stations in the RoI of the reference station are not in equal proximity, a weighting function was needed to reflect the relative importance of each station for estimation of ET_P at the reference station. The Shannon entropy was used to calculate the weight of the different geographical, climate, and statistical attributes.

2.4.1 Shannon entropy

Entropy refers to a small amount of disturbance of the thermodynamic system, and was used by Shannon (1948) to describe uncertainty in information sources. In information theory, entropy is specified as the amount of irregularity in a system. Therefore, measured entropy can be used to estimate the heterogeneity of the attributes required in ET_P estimation. The more dispersion in the amount of entropy in an attribute, the more critical it will be. The process of calculating the Shannon entropy can be expressed in a series of steps (Shannon 1948):

SE1: Normalize the decision matrix:

$$f_{ij} = \frac{x_{ij}}{\sum_{j=1}^m x_{ij}}, \quad (j = 1, \dots, m, \quad i = 1, \dots, n) \quad (2)$$

SE2: Compute entropy:

$$E_i = -k \sum_{j=1}^m f_{ij} \ln f_{ij} \quad (3)$$

$$k = \frac{1}{\ln n} \quad (4)$$

where x_{ij} is the rating of station i concerning attribute j , f_{ij} is the normalized x_{ij} , m is the number of attributes, n is the number of stations, E_i is the amount of dispersion or entropy in attribute I , and k is the entropy constant.

SE3: Determine uncertainty:

$$d_i = 1 - E_i \quad (5)$$

where d_i represents the uncertainty or degree of deviation of the data for attribute i .

SE4: Determine the significance of attribute i :

$$\widehat{W}_j = \frac{d_i}{\sum_{i=1}^m d_i} \quad (6)$$

where \widehat{W}_j denotes the attribute j weight vector.

2.5 Distance metric

In metric space, the similarity is defined by the distance criterion. If the attributes of the catchment area are the same, the measurement distance is zero. As the difference in attributes increases, the measurement distance will increase. Several methods have been proposed for determining the distance metric to express similarity, including Manhattan, Canberra, and Minkowski. In the RoI procedure, Euclidean distance is most widely used in regionalization methods. Euclidean distance is the straight line between two stations and is defined as:

$$D_{ij} = \left(\sum_{m=1}^M W_m (X_m^i - X_m^j)^2 \right)^{1/2} \quad (7)$$

where D_{ij} is the weighted Euclidean distance between stations i and j , W_m refers to the weight values of the m th attribute for the reference station that satisfy $W_m \geq 0$ and $\sum_{i=1}^m W_i = 1$, X_m^i and X_m^j are the value of the m th attribute at stations i and j , and M is the number of attributes. The distance metric matrix D is symmetrical ($D_{ij} = D_{ji}$) with zero values on its main diagonal ($D_{ii} = 0$). Since the selected attributes may have different units, it is necessary to convert the initial data before computing D_{ij} .

The most straightforward alternative is to standardize variables. In Eq. (7), X_m and Y_m are the standardized values of attributes for the reference stations.

2.6 Definition of threshold

After selecting the appropriate attributes and calculating the distance metric matrix, the first step in the RoI approach was to select a threshold value or cutoff point for the reference stations. In determining the threshold of the metric distance of the i th station, only the stations with metric distance below the threshold value will fall within the RoI of the reference station i :

$$\text{RoI} = \{j : D_{ij} \leq \theta_i\} \quad (8)$$

where RoI is a set of stations i in the region of influence and θ_i is the threshold value for station i (Burn 1990b).

Burn (1990b) presented a general framework for determining the threshold distance θ_i considering the weight of the attributes η_{ij} in three different options (#1–#3).

2.6.1 Option #1

In option #1, the RoI for the reference stations contains a limited number of stations, and all selected stations are assigned a weight within the range 0–1, expressed as follows:

$$\theta_i = \theta_L \quad \text{if} \quad \text{NSi} \geq \text{NST}, \quad (9)$$

and

$$\theta_i = \theta_L + (\theta_U - \theta_L) \left(\frac{\text{NST} - \text{NSi}}{\text{NST}} \right) \quad \text{if} \quad \text{NSi} < \text{NST} \quad (10)$$

where θ_L and θ_U are the lower and upper threshold values for station i (25th and 75th percentile of Euclidean distance), respectively, NST is the number of stations that can be nearby in RoI _{i} , and NS i is the number of stations in the RoI of the reference station. The weighting function for option #1 is:

$$\eta_{ij} = 1 - \left(\frac{D_{ij}}{\text{TP}} \right)^n \quad (11)$$

where η_{ij} is the weight of station j in the RoI _{i} , TP is the 85th percentile of the Euclidean distance for option #1, and $n = 2.5$.

2.6.2 Option #2

In option #2, a large number of stations are in the RoI of the reference station, and lower weights are allocated to stations with less similarity. In this case, the threshold value is considered:

$$\theta_i = \theta_U \quad (12)$$

The weighting function for option #2 is defined as:

$$\eta_{ij} = 1 \quad \text{if} \quad D_{ij} \leq \theta_L \tag{13}$$

and

$$\eta_{ij} = 1 - \left(\frac{D_{ij} - \theta_L}{TN - \theta_L} \right)^n \quad \text{if} \quad \theta_L < D_{ij} \leq \theta_U \tag{14}$$

In option #2, in addition to θ_L and θ_U as a weighting function, there are two other parameters (TN and n). TN is calculated using TPP as:

$$TN = \max \left[\max_{\{j\}} (D_{ij}), TPP \right] \tag{15}$$

In this case, θ_L , θ_U , and TPP are considered the 25th, 75th, and 85th percentiles of Euclidean distance and $n = 0.1$ (Burn 1990b).

2.6.3 Option #3

Option #3 is almost the same as option #2, except that all stations in the RoI of reference stations have an appropriate value of the weighting function:

$$\theta_i = \max_{\{j\}} (D_{ij}) \tag{16}$$

The weighting function for option #3 is the same as for option #2.

2.7 Clustering method

One of the agglomerative clustering methods used in this study was the Ward method. The Ward algorithm acts to minimize the internal variance of the whole cluster, by aiming to find spherical and dense clusters. It is defined thus:

$$W = \sum_{k=1}^K \sum_{j=1}^m \sum_{i=1}^{N_k} (f_{ij}^k - f_{\cdot j}^k)^2 \tag{17}$$

$$f_{\cdot j}^k = \frac{\sum_{i=1}^{N_k} f_{ij}^k}{N_k} \tag{18}$$

where W represents the total within-group sum of squares, k is the number of clusters, m is the number of attributes, N_k is the number of an attribute in stations of each cluster, f_{ij}^k is the normalized value of a j th attribute in the i th station belonging to cluster k , and $f_{\cdot j}^k$ denotes the mean value of a j th attribute for cluster k .

2.8 Regional homogeneity

The identification of homogeneous regions leads to more accurate data transfer. Homogeneous areas include stations that

are in the same group. The stations within a group have similar characteristics and, in the formation and integration of a group, all stations with similar characteristics are involved. In this regard, Hosking and Wallis (1993) evaluated several quantifications and developed the heterogeneity (H) and discordancy (D_i) measures.

The heterogeneity test is recommended to identify homogeneous regions created by regionalization. If $H < 1$, an area is considered similar; for $1 < H < 2$, a region is considered relatively heterogeneous; and for $H > 2$, a region is deemed to be heterogeneous (Hosking and Wallis 1993). The heterogeneity test contains H_1 , H_2 , and H_3 statistics, which are dependent on the L-moment distribution of linear variation coefficient (LC_V), linear skewness coefficient (LC_S), and linear kurtosis coefficient (LC_K). Hosking and Wallis found that H_2 and H_3 could not differentiate between homogeneous and heterogeneous regions and concluded that H_1 based on LC_V had the highest potential for differentiation. Therefore, H_1 is recommended as a primary index for heterogeneity and is more appropriate for this test. It is calculated as:

$$H = \frac{V - \mu_V}{\sigma_V} \tag{19}$$

where V is the weighted variance of LC_V for the studied region, μ_V is the mean of V , and σ_V is the standard deviation of V .

The test of discordancy specifies uncoordinated stations compared with the entire group in terms of the L-moment ratios. The amount of critical value for D_i (Hosking and Wallis 1997) is shown in Table 2. Stations with D_i higher than a threshold are discordant, and removing or moving discordant stations will make all regions homogeneous in the study area. The discordancy statistic is calculated as:

$$D_i = \frac{1}{3} N (\hat{u}_i - \bar{u})^T A^{-1} (\hat{u}_i - \bar{u}) \tag{20}$$

$$A = (N-1)^{-1} \sum_{i=1}^N (\hat{u}_i - \bar{u}) (\hat{u}_i - \bar{u})^T \tag{21}$$

Table 2 Critical values for the discordancy statistic (D_i)

Number of the station in the region	Critical value (D_i)
5	1.333
6	1.648
7	1.917
8	2.14
9	2.329
10	2.491
11	2.632
12	2.757
13	2.869
14	2.971
≥ 15	3

where D_i is the discordancy measure for station i , N is the number of stations in the region, \hat{u}_i is a vector containing LC_V , LC_S , and LC_K for the station, \bar{u} is the regional average for \hat{u}_i , and A is the matrix of covariance of the sample.

In regional frequency analysis, the appropriate regional distribution is considered the best fit for the stations in a homogeneous region. Therefore, the scoring method can be used to select the best regional distribution. The most commonly used goodness-of-fit methods in previous studies are the chi-square test, Kolmogorov-Smirnov test, and calculation of residual squares. The best-fit distribution can be obtained for homogeneous regions using the values of Z^{Dist} defined by Hosking and Wallis (1997):

$$Z^{Dist} = \frac{(\tau_4^{Dist} - \tau_4^R + B_4)}{\sigma_4} \tag{22}$$

where τ_4^R is an average L-kurtosis value of the region, τ_4^{Dist} is a theoretical L-kurtosis value computed from the simulation for a fitted distribution, and B_4 and σ_4 are the bias and standard deviation, respectively, of L-kurtosis values obtained from simulated data. The fitting result of the distribution is considered satisfactory if $|Z^{Dist}| \leq 1.64$. When more than one distribution qualifies for the goodness-of-fit measure, the preferred distribution is that with the lowest value (closest to zero).

3 Results and discussion

3.1 Weighting method for the defined attributes

The weight value of each attribute determines the impact that attribute will have on the desired category in determining homogeneous regions, in the present case in ULB. Analysis of the influence of climate, statistical, and geographical parameters on ETp was performed using the Shannon entropy method, and the weight of each parameter was obtained. The weights assigned to attributes in each of the categories are summarized in Table 3. Among the attributes, by far the highest weight was given to attributes belonging to the statistical group. These made up almost 73.98% of the total weight, and thus had a high degree of importance in regionalization of the basin. The climate attributes were the second most important, with 20.55% of the total weight, and finally the geographical attributes, with 5.48%.

Differences between the weights defined for each attribute within groups were observed. The range of weight changes (between the highest and lowest assigned weights) was the greatest in the statistical attribute group, indicating differences in the degree of importance of attributes in this group. The skewness of potential evapotranspiration (C_S) had the highest weight (41.6%) and was thus identified as the most influential attribute. The next most important attributes were C_V/C_S and WS, with 23.98% and 12.83% of the total weight, respectively; i.e., they also contributed strongly to regionalization.

According to the weighting results, attributes latitude and longitude (x and y), belonging to the geographical group, had the least impact, i.e., had the lowest weight (0.06%). The remaining attributes had equal influence in the reference ETp regionalization of ULB.

3.2 Weighting impact on clustering

The Ward clustering method was used to identify homogeneous regions in ULB. The hierarchical clustering algorithm in the Ward method was used to minimize the internal variance between categories. As the number of stations in each cluster decreases with increasing similarity, precise estimation of the similarity and the optimal number of clusters is required. Validation of clusters to find the optimal number of clusters was performed using the R software. The model took into account the most frequent number of clusters among 30 indicators shown in Fig. 2.

The results showed that there was no change in the number of clusters by attribute weighting, and three main clusters were created in each case for the study area. Silhouette coefficient results were used as a cluster validation index to choose the best set of clusters. The average silhouette width (ASW) is within the range -1 to $+1$, and the method with the highest ASW is optimal. In the present case, the values of this coefficient for the non-weighted and weighted clusters were 0.36 and 0.41, respectively. This indicates that attribute weighting was able to cluster the stations in ULB better than no weighting. The ASW value decreased in both cases with an increasing number of clusters, and the best clustering results were obtained at $k = 3$.

It was found that increasing or decreasing the number of attributes studied did not necessarily lead to a rise in the number of clusters. In other words, increasing or decreasing the attributes for regionalization cannot increase or

Table 3 The weight assigned to each attribute in each of the categories.

Criterion	Geographic			Climatic			Statistic		
	x	y	h	WS	RH	T	C_V	C_S	C_V/C_S
W_j	0.00057	0.00059	0.0537	0.1283	0.0359	0.0411	0.0838	0.4160	0.2398

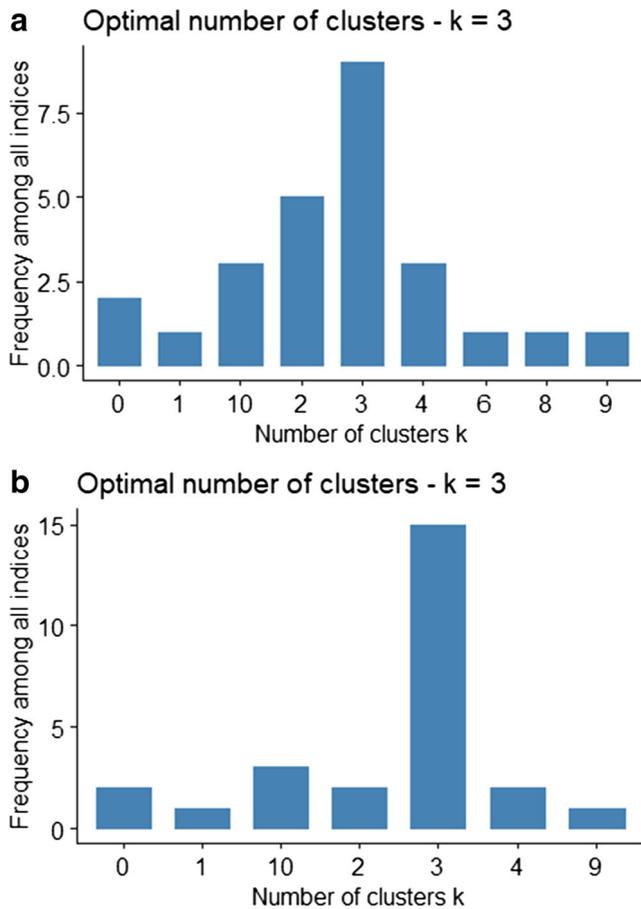


Fig. 2 The optimal number of clusters. **a** No weighting. **b** By applying weights to attributes

decrease the number of homogeneous regions in that area (Dehghan et al. 2018b).

Similarity values, obtained from Euclidean distance, affected the number of the stations in each group in clustering. Figure 3 illustrates the spatial pattern of the three homogeneous regions of ETp identified in ULB. Figure 3a shows the clustering of ETp without applying the weight, while Fig. 3b illustrates the clustering on applying weights to three categories of attributes. On comparing (a) and (b), it can be seen that clustering of the basin changed after the weighting of attributes, and that some stations were located in a different region in ULB. Thus, it can be concluded that geographical proximity is not a guarantee of similarity between stations, which is in agreement with Da Silva et al. (2017).

3.3 Regionalization with the RoI approach

The threshold values for the three reference stations (Saqqez, Tabriz, and Urmia) were determined according to the similarity distance metric. After determining the final weight of the parameters, the weighted Euclidean distance from the reference station was calculated for all stations. Considering that an increase in the metric distance between stations indicates a decrease in similarity between stations, using weighted attributes can have a positive effect on determining the metric distance between stations to enhance their similarity. The results showed that each of the three reference stations had a different threshold value than other stations, which is quite logical.

Figure 4 shows the position of the stations located in the RoI of the reference stations against the weight of each station.

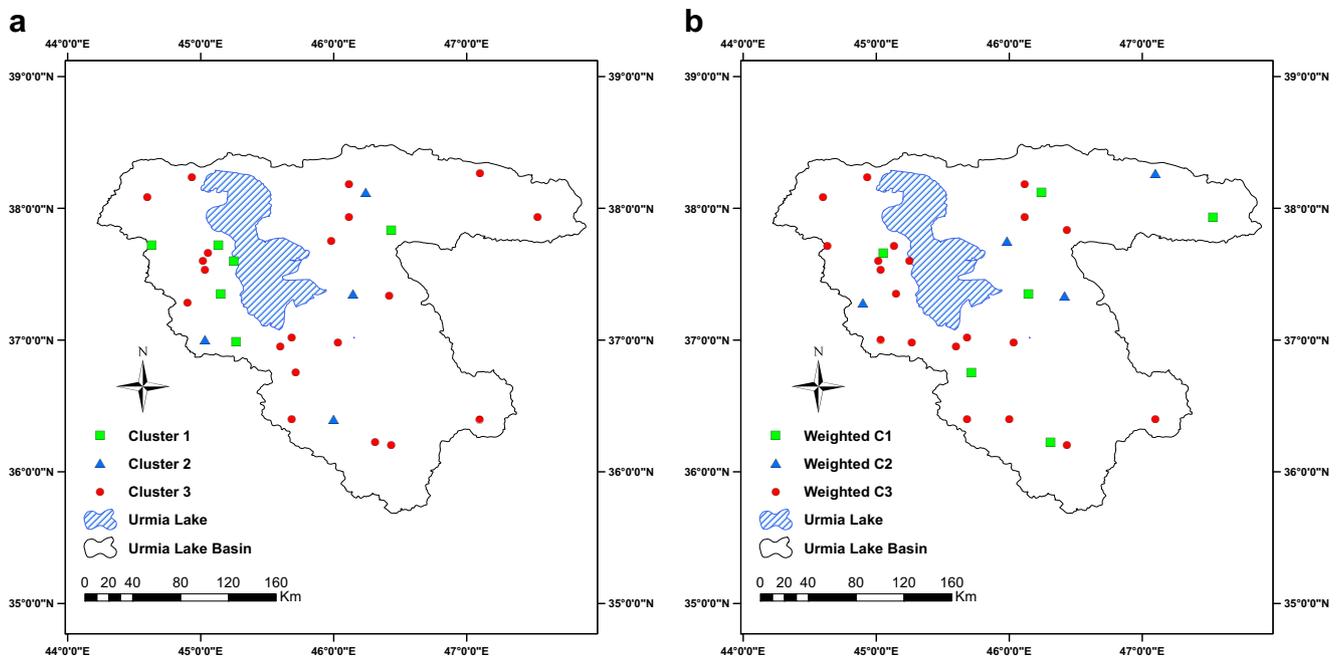


Fig. 3 Changes in the spatial distribution of ETp with the Ward clustering method in ULB. **a** Non-weighting. **b** By applying weight to attributes

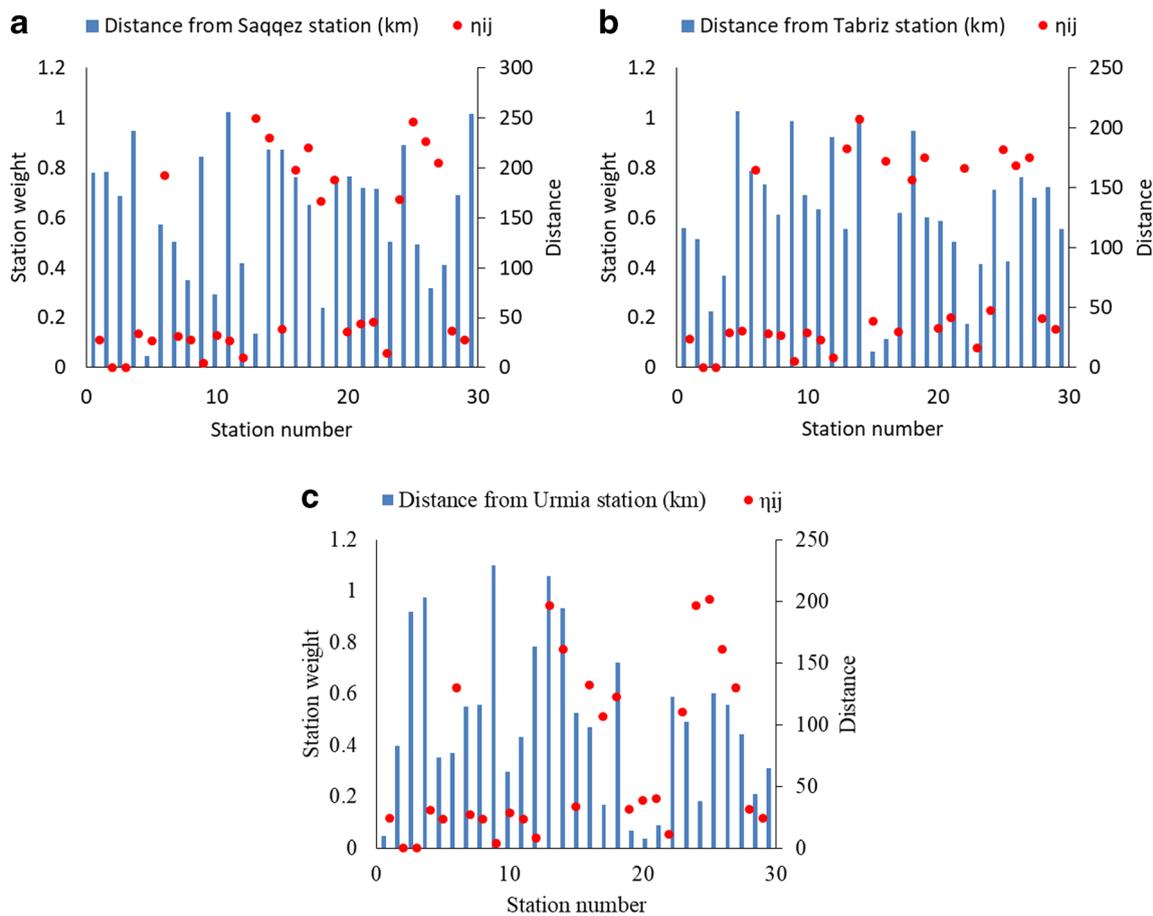


Fig. 4 Relationship between the weight of each station (η_{ij}) and its distance (km) from the reference stations: Saqqez (a), Tabriz (b), and Urmia (c) in the RoI method

As can be seen, the highest weights were assigned to stations with different distances, and some stations with different weights were near to each other. Therefore, stations closer to the reference station did not have higher weights than more distant stations.

As can be seen in Fig. 4, the stations with high weight in the RoI were located at a distance of less than 150 km from Tabriz station, about 200 km away from Saqqez station and 100 km or less from Urmia station. In general, within a distance of approximately 0 to 200 km from the reference station, weights of 0.66 to 0.99, 0.75 to 0.99, and 0.51 to 0.97 were allocated to the stations in the RoI of Saqqez, Tabriz, and Urmia stations, respectively. Thus, distance from, or proximity to, the reference station was not the most critical factor affecting the allocated weight. Closer stations to the reference station were mostly assigned higher weights, but some stations at greater distance from the reference station also had high weights. These results of the RoI approach are in agreement with previous findings (Eslamian 2010a, 2010b). Estimation of hydrological parameters in ungauged stations or station with incomplete data requires more accurate and reliable methods, such as the RoI approach. To our knowledge, this is the first study ever to estimate ETp with the RoI approach, although it has been used in flood frequency analysis (Burn

1990a, 1990b), flood regionalization (Eng et al. 2007), and precipitation frequency analysis (Gaál et al. 2008a, 2008b; Gaál and Kysely 2009; Dehghan et al. 2018a, 2018b).

3.4 Allocation of homogeneous regions

The homogeneity index was evaluated using the Monte Carlo simulation with 1000 replications for each of the areas. After calculating L-moments of LC_V , LC_S , and LC_K at each station, the discordancy (D_i) and heterogeneity (H) statistics were calculated for stations located in each area.

Stations with a high amount of D_i were removed from the set of stations to determine the homogeneous regions according to the D_i amount based on Table 2. The values of the H -statistic for homogeneous areas are shown in Table 4. Based on these values, in both cases (before and after weighting) no station was deleted in the first and second clusters, but one and three stations were detected in the third cluster before and after weighting, respectively, and were excluded from the calculations. As can be seen in Table 4, in the RoI approach, one and two stations were removed from RoI2 and RoI3, respectively, of both Urmia and Saqqez stations. One station from RoI1 and two stations from RoI3 of Tabriz station were removed.

Table 4 Values of the H-statistic of the heterogeneity test in the clustering and RoI method.

Regions	Number of Station	Removed stations with D_i test	H_1 value	
Clustering Method	No weighting			
	Cluster1	6	No	1.38
	Cluster2	4	No	0.66
	Cluster3	19	1	1.48
	By applying weight			
	Cluster1	6	No	1.27
	Cluster2	4	No	0.20
	Cluster3	17	3	1.22
	RoI Method	Urmia		
RoI1		10	No	0.88
RoI2		24	1	1.18
RoI3		30	2	1.41
Tabriz				
RoI1		10	1	0.81
RoI2		22	No	0.68
RoI3		30	2	1.05
Saqqez				
RoI1		9	No	0.86
RoI2		24	1	1.26
RoI3		30	2	1.45

After removing discordant stations, a heterogeneity test was conducted for the remaining stations in each region. According to the results of the clustering method in both cases (non-weighted and weighted based on the H_1 measure), cluster 1 and cluster 2 can be considered homogeneous regions. However, in cluster 3, H_1 exceeded the critical value of 1 representing a relatively heterogeneous region. The values of H_1 in clusters with attribute weighting were less than those in non-weighting, which indicates that the homogeneity of clusters was increased by attribute weighting.

In the RoI approach, any increase in the threshold value leads to an increase in the number of stations in the RoI of the reference station. Therefore, the number of stations was the highest in RoI3, with the highest threshold values. The lowest number of stations (9–10 stations) in the RoI of the three reference stations was observed in option #1. In options #2 and #3, 22–24 and 30 stations, respectively, were considered in the RoI of the reference stations. Thus, increasing the threshold, and thereby the number of stations in the RoI of the reference station, led to an increase in heterogeneity in the region. For the Urmia and Saqqez reference stations, the best homogeneity was observed in the area with the option #1 threshold. For Tabriz station, the homogeneity was greater with the option #2 threshold than with the option #1 threshold.

After analyzing the homogeneity of the study regions, the best-fitted distribution of these regions was determined. For

this purpose, the Z^{Dist} value for each area, including generalized logistic (GLOG), generalized extreme-value (GEV), generalized normal (LOGN), Pearson type III (P-III), and generalized Pareto (GPA), were computed (Table 5). To avoid multiple distribution functions in the estimates obtained in hydrological studies, a type of distribution function should be used for all study regions. Here, GEV, GLOG, and LOGN were determined as best-fitted distributions by the RoI approach, and GEV and GLOG by clustering in ULB. Hence, the distribution function GEV was identified as the best distribution and can be considered the selected function in all regions with both of these methods.

Root mean square error (RMSE) was used to estimate the error between simulated values and observations. In the clustering method, regions with weighted attributes had lower RMSE in comparison with non-weighted, and the highest estimated error occurred in the group with the highest number of clusters (Table 5). Based on the results obtained using the RoI approach, option #1 for the threshold showed the best performance (in terms of RMSE). Better results in terms of RMSE were obtained with the RoI method than with clustering. Unlike in the clustering method, in the RoI approach, the error values in the groups are not wide-ranging about each other and vary within a relatively low range.

4 Conclusions

In this study, regionalization of potential ETp with an integrated spatial pattern based on clustering and on the RoI approach was applied to ULB. Due to the importance of selecting attributes in regionalization, nine attributes in three groups (statistical, climate, geographical) affecting ETp were studied and weighted using the Shannon entropy method. The results showed that different attributes were allocated different weights, reflecting differences in their degree of importance. The most significant impact of weighting was found to be assigned to statistical attributes, among which skewness coefficient was identified as the most critical attribute. Thus, it can be concluded that outliers should be given special attention, as they increase the skewness coefficient.

The clustering analysis revealed differences in the clusters formed on taking into account the attributes of the study area compared with considering the conditions regardless of the attributes. Urmia Lake Basin was divided into three homogeneous regions based on cluster analysis of the study region and homogeneity tests. The optimal number of clusters was identified based on the most frequent number of clusters among 30 indicators. Average silhouette coefficient (ASW) results indicated a better performance in Ward clustering of the model with weighted attributes, in comparison with the non-weighted model. Performing a heterogeneity test and removing discordant stations increased the value of H_1 , and the

Table 5 Values of the Z^{DIST} statistic of the goodness-of-fit test for considered probability distribution functions.

Regions		Z^{Dist}					Optimal Frequency Distribution	RMSE %
		GLOG	GEV	LOGN	P-III	GPA		
Clustering Method	No weighting							
	Cluster1	2.70	0.61	1.25	1.22	-3.02	GEV	3.12
	Cluster2	2.95	1.14	1.74	1.68	-1.93	GEV	2.61
	Cluster3	0.79	-1.90	-1.41	-1.50	-7.00	GLOG	4.36
	By applying weight							
	Cluster1	1.56	-0.04	0.21	0.13	-3.12	GEV	2.54
	Cluster2	-0.12	-0.94	-0.98	-1.18	-2.67	GLOG	2.5
	Cluster3	4.07	0.62	1.67	1.62	-5.37	GEV	3.66
	RoI Method	Urmia						
RoI1		-0.03	-2.11	-1.64	-1.66	-5.95	GLOG	0.86
RoI2		2.91	-0.12	0.55	0.51	-5.73	GEV	0.95
RoI3		3.85	0.30	1.25	1.23	-6.07	GEV	2.17
Tabriz								
RoI1		1.57	-0.67	-0.05	-0.07	-4.67	LOGN	0.97
RoI2		0.25	-2.14	-1.60	-1.64	-6.55	GLOG	1.06
RoI3		3.15	-0.30	0.62	0.60	-6.49	GEV	1.91
Saqqez								
RoI1		0.04	-2.14	-1.58	-1.60	-6.07	GLOG	1.03
RoI2		2.97	-0.14	0.60	0.57	-5.82	GEV	1.23
RoI3		3.03	-0.43	0.48	0.45	-6.64	GEV	3.17

amount of H_1 in weighted attributes was better than in the non-weighted option. It can be said that attribute weighting improved homogeneity compared with when no weight was assigned to attributes of ULB.

The highest RMSE values were observed in groups with high H -statistics. Option #1 of the threshold gave the best performance (in terms of RMSE). It can be concluded that weighting of attributes in regionalization has a significant impact in obtaining accurate and reliable quantiles. One of the most important reasons for the superior performance of the RoI approach compared with clustering was the weighting of the stations, which had a significant effect in lowering the error in the estimates. Weighting the stations also reduced the role of nearby stations with low similarity to the target station. Due to coordination of the target station with other stations, the RoI method provides more accurate estimates than other regionalization methods and is a highly flexible method for transmitting information from nearby stations to target stations.

In general, the results show that the RoI is a powerful approach that rationally involves a large number of stations in the proximity of reference station, with the weight assigned to each station reflecting the lack of similarity between them. In other regionalization methods, the stations have equal weight and their relative role in the regionalization is not determined, which is one of the strengths of the RoI approach.

Acknowledgments The authors gratefully acknowledge the Water Resources Management Company for providing meteorological data from hydrometric stations and the Meteorological Organization of Iran for providing meteorological data from synoptic stations.

Funding Information Open access funding provided by University of Oulu including Oulu University Hospital.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Acreman M, Wiltshire S (1987) Identification of regions for regional flood frequency analysis. *Eos* 68:1262
- Akbari M, Torabi Haghighi A, Aghayi MM, Javadian M, Tajrishy M, Klove B (2019) Assimilation of satellite-based data for hydrological mapping of precipitation and direct runoff coefficient for the Lake Urmia Basin in Iran. *Water* 11:1624

- Al-Hamdan O, Cruise J (2009) Soil moisture profile development from surface observations by principle of maximum entropy. *J Hydrol Eng* 15:327–337
- Allen RG, Pereira LS, Raes D, Smith M (1998) Crop evapotranspiration-guidelines for computing crop water requirements-FAO irrigation and drainage paper 56, vol 300. Fao, Rome, p D05109
- Barbe D, Cruise J, Singh V (1994) Derivation of a distribution for the piezometric head in groundwater flow using entropy. In: Stochastic and statistical methods in hydrology and environmental engineering. Springer, Dordrecht, pp 151–161
- Baumgartner A, Reichel E (1975) The world water balance; mean annual global, continental and maritime precipitation, evaporation and runoff
- Bharath R, Srinivas V (2015) Regionalization of extreme rainfall in India. *Int J Climatol* 35:1142–1156
- Burn DH (1990a) Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resour Res* 26:2257–2265
- Burn DH (1990b) An appraisal of the “region of influence” approach to flood frequency analysis. *Hydrol Sci J* 35:149–165
- Castellarin A, Burn D, Brath A (2001) Assessing the effectiveness of hydrological similarity measures for flood frequency analysis. *J Hydrol* 241:270–285
- Chiu CL, Hsu SM, Tung NC (2005) Efficient methods of discharge measurements in rivers and streams based on the probability concept. *Hydrol Processes Int J* 19:3935–3946
- Chiu CL, Jin W, Chen YC (2000) Mathematical models of distribution of sediment concentration. *J Hydraul Eng* 126:16–23
- Chiu CL, Tung NC (2002) Maximum velocity and regularities in open-channel flow. *J Hydraul Eng* 128:390–398
- Da Silva HJ, Gonçalves WA, Bezerra BG (2017) Sensitivity analysis and regionalization of reference evapotranspiration for the Amazon region. *J Hyperspectral Rem Sens V* 7:258–271
- Dehghan Z, Eslamian SS, Fathian F, Modarres R (2018a) Regional frequency analysis with development of region-of-influence approach for maximum 24-h rainfall (case study: Urmia Lake Basin, Iran). *Theor Appl Climatol* 136:1483–1494
- Dehghan Z, Eslamian SS, Modarres R (2018b) Spatial clustering of maximum 24-h rainfall over Urmia Lake Basin by new weighting approaches. *Int J Climatol* 38:2298–2313
- Eng K, Milly P, Tasker GD (2007) Flood regionalization: a hybrid geographic and predictor-variable region-of-influence regression method. *J Hydrol Eng* 12:585–591
- Eslamian S (2010a) Flood regionalization using a modified region of influence approach. *JFE* 1:51–66
- Eslamian S (2010b) The physically-statistically based region of influence approach for flood regionalization. *JFE* 1:149–158
- Farina G, Alvisi S, Franchini M, Moramarco T (2014) Three methods for estimating the entropy parameter M based on a decreasing number of velocity measurements in a river cross-section. *Entropy* 16:2512–2529
- Fazel N, Haghighi AT, Kløve B (2017) Analysis of land use and climate change impacts by comparing river flow records for headwaters and lowland reaches. *Glob Planet Chang* 158:47–56
- Gaál L, Kysely J (2009) Regional frequency analysis of heavy precipitation in the Czech Republic by improved region-of-influence method. *Hydrol Earth Syst Sci* 6:273–317
- Gaál L, Kysely J, Szolgay J (2008a) Region-of-influence approach to a frequency analysis of heavy precipitation in Slovakia. *Hydrol Earth Syst Sci Discuss* 12:825–839
- Gaál L, Szolgay J, Lapin M (2008b) Regional frequency analysis of heavy precipitation totals in the High Tatras region in Slovakia for flood risk estimation. *Contrib Geophys Geodesy* 38:327–355
- Haghighi AT, Fazel N, Hekmatzadeh AA, Kløve B (2018) Analysis of effective environmental flow release strategies for Lake Urmia restoration. *Water Resour Manag* 32:3595–3609
- Holmes M, Young A, Gustard A, Grew R (2002) A region of influence approach to predicting flow duration curves within ungauged catchments. *Hydrol Earth Syst Sci* 6:721–731
- Hosking J, Wallis J (1993) Some statistics useful in regional frequency analysis. *Water Resour Res* 29:271–281
- Hosking J, Wallis J (1997) Regional frequency analysis: an approach based on l-moments. Cambridge University, Cambridge
- Jowitt P (1991) A maximum entropy view of probability-distributed catchment models. *Hydrol Sci J* 36:123–134
- Kamali MI, Nazari R, Faridhosseini A, Ansari H, Eslamian S (2015) The determination of reference evapotranspiration for spatial distribution mapping using geostatistics. *Water Resour Manag* 29:3929–3940
- Lu X, Bai H, Mu X (2016) Explaining the evaporation paradox in Jiangxi Province of China: spatial distribution and temporal trends in potential evapotranspiration of Jiangxi Province from 1961 to 2013. *ISWCR* 4:45–51
- McKenney MS, Rosenberg NJ (1993) Sensitivity of some potential evapotranspiration estimation methods to climate change. *Agric For Meteorol* 64:81–110
- Mardikis M, Kalivas D, Kollias V (2005) Comparison of interpolation methods for the prediction of reference evapotranspiration—an application in Greece. *Water Resour Manag* 19:251–278
- Merz R, Blöschl G (2005) Flood frequency regionalisation—spatial proximity vs. catchment attributes. *J Hydrol* 302:283–306
- Merz R, Blöschl G, Parajka JD (2006) Regionalization methods in rainfall-runoff modelling using large catchment samples. *IAHS* 307:117–125
- Moramarco T, Saltalippi C, Singh VP (2004) Estimation of mean velocity in natural channels based on Chiu’s velocity distribution equation. *J Hydrol Eng* 9:42–50
- Moramarco T, Singh VP (2001) Simple method for relating local stage and remote discharge. *J Hydrol Eng* 6:78–81
- Nam W-H, Hong E-M, Choi J-Y (2015) Has climate change already affected the spatial distribution and temporal trends of reference evapotranspiration in South Korea? *Agric Water Manag* 150:129–138
- Ramos J, Pelczar I, Villareal FG (2008) Variation of evapotranspiration in the Northwest of Mexico and its effect on the climate change. In: IGARSS 2008-2008 IEEE International Geoscience and Remote Sensing Symposium. IEEE, Piscataway, pp IV-635–IV-638
- Samuel J, Coulibaly P, Metcalfe RA (2011) Estimation of continuous streamflow in Ontario ungauged basins: comparison of regionalization methods. *J Hydrol Eng* 16:447–459
- Seneviratne SI et al (2006) Soil moisture memory in AGCM simulations: analysis of global land-atmosphere coupling experiment (GLACE) data. *J Hydrometeorol* 7:1090–1112
- Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423
- Shukla J, Mintz Y (1982) Influence of land-surface evapotranspiration on the earth’s climate. *Sci* 215:1498–1501
- Singh VP (2010a) Entropy theory for derivation of infiltration equations. *Water Resour Res* 46, W03527. <https://doi.org/10.1029/2009WR008193>
- Singh VP (2010b) Entropy theory for movement of moisture in soils. *Water Resour Res* 46, W03516. <https://doi.org/10.1029/2009WR008288>. Accessed 13 March 2010
- Singh VP (2011) Hydrologic synthesis using entropy theory. *J Hydrol Eng* 16:421–433
- Tasker GD, Hodge SA, Barks CS (1996) Region of influence regression for estimating the 50-year flood at ungauged sites. *JAWRA* 32:163–170
- Tsang Y-P, Felton GK, Moglen GE, Paul M (2011) Region of influence method improves macroinvertebrate predictive models in Maryland. *Ecol Model* 222:3473–3485

- Wang K, Dickinson RE (2012) A review of global terrestrial evapotranspiration: observation, modeling, climatology, and climatic variability. *Rev Geophys.* <https://doi.org/10.1029/2011RG000373>
- Wei C, Yeh H-C, Chen Y-C (2014) Spatiotemporal scaling effect on rainfall network design using entropy. *Entropy* 16:4626–4647
- Wild M, Folini D, Schär C, Loeb N, Dutton EG, König-Langlo G (2013) The global energy balance from a surface perspective. *Clim Dyn* 40: 3107–3134
- Wiltshire S (1986) Identification of homogeneous regions for flood frequency analysis. *J Hydrol* 84:287–302
- Xu CY, Gong L, Jiang T, Chen D, Singh V (2006) Analysis of spatial distribution and temporal trend of reference evapotranspiration and pan evaporation in Changjiang (Yangtze River) catchment. *J Hydrol* 327:81–93
- Zhu G, He Y, Pu T, Wang X, Jia W, Li Z, Xin H (2012) Spatial distribution and temporal trends in potential evapotranspiration over Hengduan Mountains region from 1960 to 2009. *J Geogr Sci* 22: 71–85
- Zrinji Z, Burn DH (1994) Flood frequency analysis for ungauged sites using a region of influence approach. *J Hydrol* 153:1–21
- Zveryaev II, Allan RP (2010) Summertime precipitation variability over Europe and its links to atmospheric dynamics and evaporation. *J Geophys Res Atmos* 115

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.