Regional low flow analysis in Sefidrood Drainage Basin, Iran using principal component regression

Kazem Nosrati, Gregor Laaha, Seyed Ali Sharifnia and Mojtaba Rahimi

Kazem Nosrati & *Seyed Ali Sharifnia*

Department of Physical Geography, Faculty of Earth Sciences, Shahid Beheshti University, G. C., 1983963113 Tehran, Iran

E-mail: k_nosrati@sbu.ac.ir

Gregor Laaha

Institute of Applied Statistics and Computing, University of Natural Resources and Life Sciences, BOKU Vienna, Austria

Mojtaba Rahimi

Faculty of Geography, University of Tehran, Tehran, Iran

ABSTRACT

Knowledge of low streamflow statistics is necessary for effective water management in regions prone to extreme hydrologic events such as Iran. This study employs a data set of 23 river flow time series from Sefidrood Drainage Basin, Iran, to examine regional hydrological drought based on the low flow index 7Q10. Hierarchical agglomerative cluster analysis was used to divide the 23 gauging stations into two homogeneous drought regions based on the similarity of the binary drought series. 7Q10 was determined using log-Pearson type-III and 2-parameter log-normal distributions selected as the best regional probability distribution functions in homogeneous drought region 1 and 2, respectively. The 7Q10 was related to principal components of catchment characteristics in each homogeneous drought region separately using backward stepwise regression. The resulting regression equations exhibit a coefficient of determination of 69% and 89%, respectively. The regression parameters are linked to a size factor related to catchment area, an elevation factor which is independent of catchment area, and geological formation variables, which can therefore be interpreted as important controls of low flow generation processes in the study area. The equations developed here are expected to provide robust estimates of 7Q10 values for watersheds in areas of similar geomorphology, geology and climate.

Key words | base flow index, cluster analysis, hydrological drought, principal component analysis, Sefidrood Drainage Basin

INTRODUCTION

Droughts have serious impacts on ecosystems and society. In 1998–2001 the extended drought affected more than 96% of the area of Iran by water-use restrictions, land subsidence and forest fires (Agrawala *et al*. 2001; Tabrizi *et al*. 2010). Hydrological drought occurs when river streamflow and water storages in aquifers, lakes or reservoirs fall below long-term mean levels and often affects large areas. Several factors, including lack of, or less frequent precipitation, poor water management and erosion, can cause or enhance hydrological drought (Dai 2011). Knowledge about hydrological droughts is important for a variety of tasks, e.g., water quality management, determination of minimum downstream flow requirement for hydropower and ecological needs, irrigation system design and wastewater treatment. Thus, prediction of hydrological drought indices and its regional variability is essential for a sustainable management of water resources.

Hydrological droughts can be assessed in terms of streamflow quantity by means of low flow indices, such as flow quantiles or the lowest annual flow for a given duration (e.g., 1, 7, 15 and 30 days). For seasonal climates, it has been recommended to calculate low flow indices for each season separately (Tallaksen *et al*. 1997; Laaha & Blöschl 2006b). A comprehensive review of the low flow characteristics and methodologies has been given by Smakhtin (2001). In streamflow drought studies, drought events can also be characterized by indices derived by the threshold level method (Tallaksen & van Lanen 2004), such as mean or distribution of drought durations and deficit volumes. The threshold level specifies some statistically optimal or purpose-related statistic of the drought variable and serves to divide the time series into deficit and surplus sections.

The frequency and severity of hydrological drought is often defined on a watershed or river basin scale (NDMC 2005). When a regional assessment is sought, the low flow characteristics of sites where no streamflow measurements are available need to be estimated by means of regionalization. For this purpose, regression models are frequently used to relate low streamflow indices to physical catchment characteristics describing climatic, topological, geological and other properties of the catchment. The established relationships can subsequently be applied to predict low flows at ungauged sites in the entire region (Smakhtin 2001). Regression models have become a standard tool for prediction of low flow statistics (e.g., 7Q10, the annual minimum 7-day average streamflow occurring once every 10 years on average) at ungauged sites and have been widely reported in the literature (e.g., Vogel & Kroll 1992; Dingman & Lawlor 1995; Brandes *et al*. 2005; Hejazi & Moglen 2007).

In a heterogeneous area, the linear relationships of the stream flow index with catchment characteristics may vary between the regions, and regression equations based on hydrologically homogeneous subregions may be used with greater confidence to predict low flow (Nathan & McMahon 1990). This is often termed regional regression approach. Laaha & Blöschl (2006a, 2006b) assessed the value of catchment grouping for low flow regionalization and found a gain in performance depending on the grouping method. For the Austrian study area, classifications based on seasonality measures clearly outperformed the global regression model. Many studies, therefore, have investigated the delineation of a homogeneous region by using different sets of clustering variables, such as the statistical properties of drought events (Stahl & Demuth 1999; Fleig *et al*. 2011; Hannaford *et al*. 2011; Nosrati 2012) and the catchment characteristics including climatic, geomorphological and soil variables (Nathan & McMahon 1990; Rifai *et al*. 2000; Yu *et al*. 2002; Nosrati *et al*. 2004; Laaha & Blöschl 2006a; Nosrati & Shahbazi 2008; Mamun *et al*. 2010).

Iran lies approximately between 40°E and 64°E in longitude and 25°N and 40°N in latitude. The topographic features of Iran show two major mountain ranges, Alborz and Zagros, in the north and west of Iran, respectively, which surround the arid and semi-arid region of the central part of Iran. The mean annual precipitation of Iran varies between 1,800 mm in the north to less than 100 mm in the central arid regions of the country. The precipitation coefficient of variation (CV) in stations varies between 20% and 75% from wet to dry regions of Iran. Surface water is not only a major source of drinking water in Iran, but also supplies public water utilities and accounts for almost all of the water supply to rural households. In Iran, the availability of water resources is critical during certain periods. River flows are strongly seasonal, characterized by low natural flow during summer. The high frequency of droughts makes it necessary to improve management strategies for water quantity and quality during dry periods. Therefore, understanding and predicting low flows is important for decision-makers involved in water resources management to account for the availability of water supply, the quality and quantity of water for human use, recreation, or irrigation purposes and wildlife conservation, especially in areas prone to extreme hydrological events.

A regional low flow analysis provides a valuable framework for assembling the information necessary for understanding and predicting low flow and drought at ungauged sites, or at sites where data are incomplete. Establishing a regional low flow analysis provides a means of clarifying the link between low flow indices and catchment descriptors using regression techniques. In this study, information on the time patterns of drought occurrence is

used as a basis for regionalization (Figure 1), similar to the approach Stahl & Demuth (1999) used for linking streamflow drought to the occurrence of atmospheric precipitation patterns. Catchments are regarded as similar if streamflow drought events show similar time patterns, thus occur simultaneously. A cluster analysis based on binary drought occurrence series is applied to find homogeneous drought groups or regions as a basis for regionalization, and the homogeneity by two different streamflow measures are checked. Further, regionalization is performed in homogeneous drought regions by principal component regression. Thus, the stepwise multiple regressions were used based on catchment characteristics and low flow to develop a reliable and statistical-based method for assessing the low flows in homogeneous drought regions in the Sefidrood Drainage Basin, Iran.

MATERIALS AND METHODS

Study area

The study was performed in the Sefidrood Drainage Basin, located between 46° 27′ and 51° 11′ east longitude and 34° 58′ and 37° 56′ north latitude, in northwestern Iran within the boundaries of Hamadan, Kordestan, Zanjan, East Azerbaijan, Ghazvin and Guilan Provinces (Figure 2). The Sefidrood River, the main river of the area, originates in the Alborz and Zagros Mountains and, after junction with some main tributaries, flows into the Sefidrood Dam Reservoir and finally into Sefidrood River estuary and the Caspian Sea. The study area is an important region not only for agricultural production, which directly depends on river water resources, but also for hydropower water supply and irrigation schemes in downstream areas. The dense population and an increasing demand for irrigation has led to a dramatic reduction of streamflows during the last decades. The investigations also show that sedimentation in the Sefidrood Dam Reservoir and the estuary of the Sefidrood River is a significant visual impact of land degradation by human activities (Azari Dehkordi *et al*. 2003). In spite of this, a sufficient streamflow measurement network does not exist. Hence, a better understanding of quantity and regional variability of low flows could assist in the development of future flow restoration and management strategies on the study area.

The drainage area of the Sefidrood Drainage Basin is 59,273 km^2 which includes 23,000 km^2 (38.8% of total area) crop fields, 180 km^2 (0.3% of total area) residential rural area, 34,193 km^2 (57.7% of total area) natural rangelands and 1,900 km^2 (3.2% of total area) natural forest. The drainage basin has variable lithological characteristics, with outcrops of Pre-Cambrian to Quaternary formations. The Sefidrood Drainage Basin has a mountainous

topography, with a minimum and maximum height of 1,690 m and 4,407 m above the sea level in the Sefidrood Dam and Taleghan Mountains, respectively. The range of precipitation is 375 to 585 mm. Long-term series (1975–2005) mean annual precipitation (P) in the study area is strongly dependent on height (H). This relationship was explored by calculating linear regression based on mean annual precipitation and height data of 23 studied stations, yielding the following regression equation (coefficient of determination $R^2 = 0.88$):

P = 17.07 + 0.196H   (1)

Streamflow data

Natural daily discharge series, from the archives of the Water Resources Research Organization, Iran were obtained for 23 stations in the region. Selected river gauges with a continuous 10-year record (1996–2006) were used in this study. The periods of missing data were filled by regression against the most highly correlated station. The record length is, in fact, rather short for long period assessments of water resources, but appears to be well suited to give an accurate characterization of the current low flow situation (Laaha & Blöschl 2005).

Catchment characteristics

For each catchment drainage area, perimeter, mean slope gradient, mean, maximum and minimum elevation, main stream length, summation of stream lengths, drainage density, length and width of rectangle-equivalent, circularity ratio, time of concentration, mean annual precipitation, urban and forest fraction per cent, and percentage of geological formations with high, medium, low and very low infiltration capacity were determined (Table 1).

Homogeneous drought regions

Binary daily series of drought and non-drought days were derived using streamflow deficiency indicator (DI) defined by a varying threshold level method. Streamflow drought was defined by the flow that is exceeded 90% of the time (the Q90 flow) as threshold level which has the advantage of removing the influence of streamflow seasonality on droughts. For a given day j, the daily-varying Q90 value is calculated by ranking all historical values on day j plus 15 days either side of day j. The window either side of the day of interest helps increase the size of the sample, and gives a smoother flow duration curve (FDC) than would result from just one value per year of day j (Hannaford et al. 2011). The choice of a percentile from the FDC as threshold level depends on hydrological regime. A range of thresholds from Q70 to Q95 is considered reasonable for perennial streams (Tallaksen & van Lanen 2004). Hannaford et al. (2011) and Stahl (2001) used Q90 as reasonable threshold level across Europe, whereas Fleig et al. (2011) found Q70 to be reasonable threshold level in Denmark and Great Britain. This approach is described in more detail in Stahl (2001), Tallaksen & van

Lanen (2004) and Hannaford *et al* (2011). For a given day j, the DI was calculated by the following equation:

DI(j) = 1 if Q(j) ≤ Q90(j)

DI(j) = 0 if Q(j) > Q90(j)   (2)

Homogeneous drought regions were identified by cluster analysis (CA) on the binary drought series. Hierarchical agglomerative CA was performed on the data set by means of Ward's method, using Euclidean distances as a measure of similarity. After calculating Euclidean distance as a measure of similarity, a linkage or amalgamation rule is needed to determine when two clusters are sufficiently similar to be linked together. The Ward's method attempts to minimize the sum of squares of any two clusters that can be formed at each step. The spatial variability of selected river gauges based on the binary drought series was determined using the linkage distance, reported as D_{link}/D_{max}, which represents the quotient between the linkage distances (D_{link}) for a particular case divided by the maximal linkage distance (D_{max}). The quotient is then multiplied by 100 as a way to standardize the linkage distance represented on the x-axis of the dendrogram.

In order to compare the hydrological characteristics within and between the identified homogenous drought regions, basic streamflow characteristics including the mean specific discharge (q, mean daily discharge divided by basin area) and the base flow index (BFI), were calculated for each station based on the daily streamflow series.

Regional probability distribution functions

The low flow analyses presented in this paper focus on the low flow characteristic 7Q10. This low flow characteristic has been chosen because of its relevance for numerous water resources management tasks in the study area, including water quality control, river ecology and environmental flow management (e.g., Smakhtin 2001; Dudley 2004; Eslamian *et al*. 2010; Mamun *et al*. 2010), and assessment of hydrological change (Ryu *et al*. 2011). In order to determine the suitable regional probability distribution that optimally fits the minimum 7-day low flow values in the homogeneous drought regions, first, the annual 7-day minimum discharge series for each gauge were computed. Then, ten probability distributions including the normal, 2-parameter log-normal (LN2), 3-parameter log-normal (LN3), gamma, Pearson type-III (PIII), log-Pearson type-III (LPIII), generalized logistic (GLOG), generalized extreme value (GEV), 3-parameter Weibull (W3), generalized Pareto (GPAR) distributions were evaluated to determine which distribution most appropriately fit the low flow data. The Kolmogorov–Smirnov test and ranking method were used to determine the best fitting distributions in the homogeneous drought regions using EasyFit 5.5 (MathWave

Technologies 2010). There are a number of well-known methods including method of moments, maximum likelihood estimates, least squares estimates and method of L-moments which can be employed to estimate distribution parameters. For every supported distribution, EasyFit implements one of the parameter estimation methods that has the best results using the Kolmogorov–Smirnov test as goodness of fit (GOF) test and significance level (0.05) as optimality criteria. In order to determine the best distribution, a ranking method was used. In this method, ten scores ranging from 1 to 10 related to ten used distributions were assigned to each gauging data set such that 1 was given to the distribution which best fitted the data, 2 to the distribution which fitted the data in second order and so on. The summation of scores shows the suitability of distribution such that the best distribution got the lowest sum of scores. The selected regional probability distribution function in each homogeneous drought region was then used to calculate the annual 7-day minimum discharge series with a 10-year return period (7Q10).

Regionalization of low flow

The regionalization of low flow indices based on catchment characteristics covers both the regional analysis of low flows and the estimation of low flow characteristics at ungauged sites. We used regional principal component regression as the regionalization method where, for each homogeneous region resulting from the cluster analysis, a principal component analysis (PCA) is applied to transform catchment characteristics into uncorrelated variables. The most significant variables are subsequently used as predictor variables in a multiple regression model. Regression analysis based on component scores ensures that the independent variables are a parsimonious subset capturing the underlying dimensions of the full set of potential independent variables, and that they are uncorrelated as well. The method is therefore well suited to fit regressions in case of multicollinearity (Rogerson 2001).

In our study, principal components with eigenvalues >1 were selected and subsequently subjected to a varimax rotation to minimize the number of variables that have high loadings on each component (Demuth 1993; Hill & Lewicki 2007). In addition, communalities of every single variable for component model were calculated to estimate the portion of variance in each variable explained by the rotated principal components. Component scores for each catchment in the homogenous regions were calculated and these components were used as independent variables in stepwise multivariate regression analyses to develop the best equations (models) able to predict 7Q10. The regression statistics including adjusted R^2 and the smallest p-value of the F-test were used to provide the best subset selection of predictors by examining all possible regressions. Multicollinearity among

the model predictors was evaluated by the variance inflation factor (VIF), using 10 as a cut-off value. All statistical analyses were performed using STATISTICA V. 8.0 (StatSoft 2008).

RESULTS AND DISCUSSION

Homogeneous drought regions

CA was used to separate the 23 stations into groups with similar binary daily drought series. It yielded a dendrogram (Figure 3(a)) grouping all stations of the study area into two statistically significant clusters at $(D_{link}/D_{max}) \times 100 < 50$. The graph of amalgamation schedule represents a line graph of the linkage distances at successive clustering steps. This graph can help in the selection of a cut-off for the dendrogram and, consequently, it helps in the determination of the optimal number of clusters (Hill & Lewicki 2007). The number of clusters was chosen because the graph of amalgamation schedule shows a sudden increase in D_{link} when combining the last two clusters (i.e., step 22; Figure 3(b)). Thirteen and ten stations were classified in cluster 1and cluster 2, respectively (Figures 2 and 3(a)).

Median, quartiles and range values of q and BFI for the two clusters were determined (Figure 4). The results of a t-test showed that q (p < 0.001) and BFI (p = 0.03) were significantly different between the regions. The high q value within cluster 2 was caused by stations with outstandingly high annual precipitation (mean annual precipitation of 624 mm, which is 300 mm greater than the average for cluster 1). The higher BFI values in cluster 2 may be related to the high fraction of forests in all catchments in cluster 2. These results conducted by Fleig *et al*. (2011) confirmed that drought regions (clusters) provide useful hydrological differentiation by the regional streamflow characteristics (q and BFI) in northwestern Europe.

It is interesting that the two clusters form contiguous regions and seem to represent similar climate, geology and land use conditions that have been shown to have significant influence on hydrological drought (e.g., Talleksen & Van Lanen 2004). They are physically divided east/west in upland/coastal regions, respectively (Figure 2). Cluster 1 describes the regions at higher altitudes up to 4,312 m.a.s.l, where low flow conditions are affected by snow storage and freezing. This gives rise to a mixed winter and summer low flow regime. Winter low flows are directly caused by freezing processes, as is typical for alpine climates. In addition, summer low flows occur, but they are less pronounced than in coastal areas, as they are fed by snow melt and groundwater sources. Cluster 2, however, contains forested, mountainous regions of lower altitude that are affected by coastal climate. Here the stations exhibit considerably less seasonal low flow regimes, but are subject to multiannual droughts.

The higher percentage of geological formations with medium and high infiltration capacity in cluster 2 increases the base flow and consequently decreases the short-term rainfall deficiency effects on low streamflow, with the base flow providing an effective buffer during dry spells. When rainfall deficits extend over longer timescales, base flow-dominated catchments are more vulnerable than more responsive catchments. This is because the prolonged reduced rainfall restricts groundwater replenishment, meaning that once rainfall does return, it is not sufficient to stimulate a recovery in river flows because groundwater levels must be restored before base flow contributions to river flow can recommence.

Regional probability distribution functions

For each drought region data series, the ranking method was performed to identify the most appropriate regional probability distribution function. The obtained scores for the selected distributions are shown in Table 2. The sum of scores for each distribution showed that log-Pearson type-III (LPIII) got the lowest value (40) for drought region (cluster) 1 and therefore was chosen as the representative distribution in this region. Pearson type-III and normal distributions got the highest values (90 and 89) and represent the worst fitting distributions (Table 2). For drought cluster 2, the 2-parameter log-normal (LN2) distribution was most often selected (got the lowest scores) and is regarded as the best fitting regional probability distribution function for drought cluster 2. The 3-parameter log-normal and Pearson type-III distributions got the highest sum of scores and fit worst (Table 2). For the whole study area, LN2 and LPIII with scores 102 and 103, respectively, were selected as the best regional distributions whereas PIII got the highest sum of scores and appears as the worst fitting distribution (Table 2).

Overall, LN2 and LPIII perform best, and it is interesting to compare this result with results of other studies conducted in the same, or in hydroclimatologically similar areas. For Atrak basin in the northeast of Iran, Nosrati *et al*. (2002) recommended LPIII and LN2 distributions for 7Q10 frequency analysis and concluded that the LPIII is the best distribution for short duration low flows such as 7-day low flows. Modarres (2008) found GLOG and GEV as parent distributions for regional low flow frequency analysis in eastern and western regions of the north of Iran, respectively. Modarres & Sarhadi (2010) selected LN3 as a regional distribution function for the extreme hydrologic drought periods in the southeastern arid region of Iran, while LN3 consistently ranks among the worst performing distributions in this study. This difference can be caused by variations in lithology and climate parameters. Durrans & Tomic (1996) concluded that the LPIII is a suitable candidate for low flow modelling in 128 gauged stations in the USA. Chen *et al*. (2006) recommend the LN3

distribution function for the south of China for regional low flow frequency analysis. Tasker (1987) recommended W3 and LPIII distributions to describe the frequency of 7-day annual low flow series for 20 rivers in Virginia. Vogel & Kroll (1989) recommended LN2, LN3, LPIII and W3 distributions for 23 sites in Massachusetts.

LN2 and LPIII were used to calculate the 7Q10 in homogeneous drought clusters 1 and 2, respectively. Median, quartiles and range values of 7Q10 for the two clusters were determined (Figure 5). The results of a t-test showed that 7Q10 ($p < 0.0001$) was significantly different between the regions.

Regionalization of hydrological drought

For each of the two drought regions resulting from CA, a separate principal component analysis was performed on the normalized data sets to identify the components replacing the most important variables. In a subsequent step, the principal components were subjected to varimax rotation in order to maximize correlations with catchment characteristics, to increase their interpretability.The results showed that for cluster 1, the first four principal components (PCs) with eigenvalues >1 accounted for 85% of variability in catchment characteristics (Table 3). Communalities for catchment characteristics indicate the four PCs explained >90% of variance in drainage area, perimeter, length of rectangle-equivalent, maximum and mean elevation, main stream length, summation of stream lengths, time of concentration, percentage of geological formations with low and very low infiltration capacity. The four PCs explained >80% of variance in minimum elevation, mean slope, mean annual precipitation and percentage of geological formations with high infiltration. The four PCs also explained >70% of variance in width of rectangle-equivalent, circularity ratio, drainage density and urban fraction per cent, and <30% in percentage of geological formations with medium infiltration capacity (Table 3). A high communality estimate suggests that a high portion of variance was explained by the component; therefore, it would get higher preference over a low communality estimate. Thus, percentage of geological formations with medium infiltration capacity was the least important attribute due to the lowest communality estimates in cluster 1, since it is not highly correlated with the four components.

For the data set of cluster 1, PC1 explained the largest proportion (50.5%) of total variance. PC1 had a strong positive loading (>0.75) on catchment drainage area, perimeter, circularity ratio, length and width of rectangle-equivalent, main stream length, summation of stream lengths, and time of concentration, a strong negative loading on mean annual precipitation and a moderate positive loading on mean slope and thus may be considered to represent a magnitude effect or size factor related to catchment area (Table 3). The negative

loading on mean annual precipitation for PC1 can be interpreted as secondary effects due to their correlation with catchment area. PC2 explained 14.3% of the total variance, and was characterized by high positive loading on percentage of geological formations with high infiltration capacity and high negative loadings on minimum elevation and percentage of geological formations with low infiltration capacity (Table 3). As the proportion of areas with high infiltration capacity decreases with catchment altitude, the relationship with minimum elevation constitutes a secondary effect, and because of these intercorrelations we interpret PC2 as a geology factor. PC3, explaining 13.1% of total variance, has strong positive loadings on maximum and mean elevation, and moderate negative loadings on drainage density and the fraction of urban areas and thus may be considered to represent the effect of elevation which is independent of catchment area. Again, we interpret the moderate loadings on urban fraction as a secondary effect, due to lower proportion of urban areas in higher altitudes. PC4, explaining 7.2% of total variance, was characterized by a high negative loading on percentage of geological formations with very low infiltration capacity which suggests that it may represent the importance of this geological formation.

For cluster 2, four components with eigenvalues >1 were identified. A summary of varimax rotated component loadings on the catchments' descriptive variables is given in Table 3. These four components explained >89% of variability in catchment characteristics in cluster 2. From the communalities presented in Table 3 it becomes clear that drainage density, urban fraction per cent and percentage of geological formations with medium infiltration capacity are the least important catchment characteristics. Investigating these characteristics in the catchments showed that urban areas have been established in geological formations with medium infiltration capacity, so these factors are obviously correlated. The same is true for drainage density which shows a similar spatial pattern as geological formations with medium infiltration capacity. These variables have only minor effects on low flows.

For the data set of cluster 2, PC1 explained 48.4% of total variance. PC1 was characterized by high positive loadings on catchment drainage area, perimeter, circularity ratio, length and width of rectangle-equivalent, main stream length, summation of stream lengths and time of concentration (Table 3), and thus may be considered, again, to represent a magnitude effect or size factor related to catchment area. There are further a negative loading on mean annual precipitation and a moderate positive loading on urban fraction and percentage of geological formations with medium infiltration capacity which we interpret as their correlation with catchment area and land use. PC2 was characterized by high positive

loading on mean slope, drainage density and forest fraction per cent, and moderate negative loadings on minimum and mean elevation (Table 3). As the proportion of areas with forest fraction per cent decreases with catchment altitude, PC2 may be considered to represent land use. PC3 was characterized by a high positive loading on maximum elevation and a high negative loading on percentage of geological formations with very low infiltration capacity and thus may be considered to represent the elevation factor which is independent of catchment area. We interpret the high negative loading on percentage of geological formations with very low infiltration capacity as a secondary effect, due to lower proportion of formations with very low infiltration capacity in higher altitudes. PC4 was characterized by a high positive loading on percentage of geological formations with low infiltration capacity and a moderate negative loading on percentage of geological formations with high infiltration capacity, which suggests that it may represent the importance of this geological formation.

In a subsequent step of the regionalization procedure, component scores were calculated for each cluster separately, using the resulting component score coefficient matrix. The resulting components were used as independent variables in stepwise multivariate regression analyses. The analysis was, again, performed for the two clusters separately. A summary of stepwise multivariate regression models examining the extent to which the extracted components explain the 7Q10 for each drought region (cluster) in the study area is given in Table 4.

The analysis of the cluster 1 data set showed that PC1 is the most significant predictor of 7Q10 low flow (Table 4). The equation fitted through multiple regression using this variable was able to explain 69% of the 7Q10 variation of the studied catchments. Mean absolute error (MAE) and root mean squared error (RMSE) of the model were 0.21 and 0.26, respectively. The observed and predicted values for the 7Q10 values of cluster 1 are plotted in Figure 6(a). Therefore, 7Q10 is influenced by size factor related to catchment area.

Alternative sets of predictors were also evaluated for cluster 2 data sets. For these data sets, PC1, PC3 and PC4 were selected as 7Q10 predictor explaining 89% of the 7Q10 variation (Table 4) of the studied catchments. MAE and RMSE of the model were 0.71 and 0.85, respectively. The predicted values for the 7Q10 are plotted against the observed values for the 7Q10 for cluster 2 in Figure 6(b). The t statistics showed that PC1 is the most important factor in the model (Table 4). Therefore, 7Q10 is influenced by a size factor related to catchment area, an elevation factor which is independent of catchment area, and geological formation variables. Based on the coefficient of regression model and precipitation factor

loadings in PC1, the negative relationship between 7Q10 and precipitation is dissimilar to other studies that showed that magnitude effect or size factor related to catchment area can control the low flow discharge and has a dominant effect compared with the precipitation. Since the annual precipitation has been related to a 7-day low flow, this anomalous negative relationship may be subject to a seasonal effect.

The 7Q10 was positively related to the proportion of geological formations with low infiltration capacity as well as negatively linked to the proportion of geological formations with high infiltration capacity in PC4. There is a different source of low flow losses which operates in karst regions in the study area. In our model, it is expected that part of the variance that cannot be explained by the independent variables and their interactions or reverse relationship is due to the fact that differences between catchments are not entirely accounted for by the characteristics that were included in the statistical analysis.

This is in agreement with findings from many previous studies of low flow regionalization that have demonstrated a significant correlation between the low flow and catchment characteristics. Nosrati & Shahbazi (2008) examined the regional low flow of 16 stations of the Atrak River (northeastern Iran) based on 17 catchment characteristics and found that drainage area, slope and percentage of permeable geological formations accounted for 92% of the spatial variability of the 7Q10 flows in the hybrid multiple regression analysis. Modarres (2008), in north Iran, found that drainage area is the main factor affecting low flow. Eslamian *et al*. (2010) showed that size and geographic position are the main factors affecting 7Q10 in Karkhe Basin, Iran. Vogel & Kroll (1990) determined that drainage area, mean annual precipitation and basin relief were significant parameters. Vogel & Kroll (1992) also found close to direct proportionality of 7Q10 with a watershed area in Massachusetts. Brandes *et al*. (2005) included the recession constant (that depended on drainage density, landscape slope, bedrock geology and soil infiltration rate) as a model parameter for 7Q10 estimation equations. Hejazi & Moglen (2007) regionalized low flow for six urbanized watersheds in the Maryland Piedmont region, resulting in a regression that included precipitation, temperature, imperviousness in the watershed and area of the watershed. Rifai *et al*. (2000) created regression equations for the 7Q10 flow for Texas based on meteorological and physiographic data from 63 gauged streams. The regression parameters included drainage area, channel slope, predominant hydrologic soil group and the precipitation. Using data from 60 gauging stations, Flynn (2003) used total drainage area, mean summer precipitation and average mean annual temperature to predict 7Q10 flows for New Hampshire streams. Dudley (2004) used 26 gauging stations on rural rivers in Maine to

develop regression equations. 7Q10 regression equation used drainage area and fraction of the drainage basin underlain by sand and gravel aquifers.

CONCLUSION

In this case study, we examined regional hydrological drought based on the low flow index 7Q10 for the Sefidrood Drainage Basin, Iran. Analyses were based on a data set of 23 daily discharge time series measured over a 10-year standard observation period. The paper used state-of-the-art methods in a novel application/region-drought in Iran. Cluster analysis techniques were applied based on binary drought occurrence series, using the flow quantile Q90 as threshold level, to find homogeneous drought regions as a basis for regionalization. The analysis yielded two clusters which show similar time patterns of drought events, differ significantly in terms of mean flow and the base flow index, and form contiguous regions in space. For each gauge, the low flow index 7Q10 was determined using log-Pearson type-III (LPIII) and 2-parameter log-normal (LN2) distribution, which were selected in a comparative analysis as the best regional probability distribution function in homogeneous drought regions 1 and 2, respectively. For regionalization of 7Q10, principal component analysis aided in extraction and identification of the most important catchment characteristics. The resulting principal components were related to 7Q10 low flows in each homogeneous drought region separately, using backward stepwise regression. The thus-obtained regression equations exhibit a coefficient of determination of 69% and 89%, respectively. The regression parameters are linked to a size factor related to catchment area, an elevation factor which is independent of catchment area, and geological formation variables. All component loadings in sign and magnitude are well interpretable on hydrological grounds, and can therefore be interpreted as important controls of low flow generation processes in the Sefidrood Drainage Basin.

Taken together, the regional equations developed by principal component regression are expected to provide estimates of hydrological drought sensitivity, and 7Q10 values for watersheds in areas of similar geomorphology, geology and climate. However, they have not yet been tested and thus require further investigation as to their practical limits. It would also be interesting to extend the analysis to a regionalization of the entire low flow distribution, in a regional frequency analysis framework. These questions will be treated in a subsequent study, and results will be reported in future publications.

REFERENCES

Agrawala, S., Barlow, M., Cullen, H. & Lyon, B. 2001 *The drought and humanitarian crisis in Central and Southwestern Asia: a climate perspective* (No. IRI special report No 01-11). International Research Institute for Climate and Society, Palisades.

Azari Dehkordi, F., Makhdoum, M. F. & Nakagoshi, N. 2003 Sefidrood river sub-watershed-dam-estuary and degradation model: A holistic approach in Iran. *Chinese Geographical Science* **13**(4), 328–333.

Brandes, D., Hoffmann, J. G. & Mangarillo, J. T. 2005 Base flow recession rates, low flows, and hydrologic features of small watersheds in Pennsylvania, USA. *JAWRA J. Am. Water Resour. Ass.* **41**(5), 1177–1186.

Chen, Y. D., Huang, G., Shao, Q. & Xu, C. Y. 2006 Regional analysis of low flow using L-moments for Dongjiang basin, South China. *Hydrolog. Sci. J.* **51**(6), 1051–1064.

Dai, A. 2011 Drought under global warming: a review. *Wiley Interdisciplinary Reviews: Climate Change* **2**(1), 45–65.

Demuth, S. 1993 *Untersuchungen zum Niedrigwasser in West-Europa.* [Investigations of low flow in southwest Germany] Freiburger Schriften zur Hydrologie, Band 1, Freiburg, Germany.

Dingman, S. L. & Lawlor, S. C. 1995 Estimating low flow quantiles from drainage-basin characteristics in New Hampshire and Vermont. *JAWRA J. Am. Water Resour. Ass.* **31**(2), 243–256.

Dudley, R. W. 2004 *Estimating Monthly, Annual, and Low 7-Day, 10-Year Streamflows for Ungaged Rivers in Maine*: US Dept. of the Interior, US Geological Survey.

Durrans, R. S. & Tomic, S. 1996 Regionalization of low-flow frequency estimates: an Alabama case study. *JAWRA J. Am. Water Resour. Ass.* **32**(1), 23–37.

Eslamian, S., Ghasemizadeh, M., Biabanaki, M. & Talebizadeh, M. 2010 A principal component regression method for estimating low flow index. *Water Resour. Manage.* **24**(11), 2553–2566.

Fleig, A. K., Tallaksen, L. M., Hisdal, H. & Hannah, D. M. 2011 Regional hydrological drought in north-western Europe: linking a new Regional Drought Area Index with weather types. *Hydrol. Process.* **25**(7), 1163–1179.

Flynn, R. H. 2003 *Stream-gaging Network Analysis for the 7-day, 10-year Annual Low Flow in New Hampshire Streams*: US Dept. of the Interior, US Geological Survey.

Hannaford, J., Lloyd-Hughes, B., Keef, C., Parry, S. & Prudhomme, C. 2011 Examining the large-scale spatial coherence of European drought using regional indicators of precipitation and streamflow deficit. *Hydrol. Process.* **25**(7), 1146–1162.

Hejazi, M. I. & Moglen, G. E. 2007 Regression-based approach to low flow prediction in the Maryland Piedmont region under joint climate and land use change. *Hydrol. Process.* **21**(14), 1793–1801.

Hill, T. & Lewicki, P. 2007 *STATISTICS: Methods and Applications*. StatSoft, Tulsa, OK.

Laaha, G. & Blöschl, G. 2005 Low flow estimates from short streamflow records—a comparison of methods. *J. Hydrol.* **306**(1–4), 264–286.

Laaha, G. & Blöschl, G. 2006a A comparison of low flow regionalisation methods—catchment grouping. *J. Hydrol.* **323**(1), 193–214.

Laaha, G. & Blöschl, G. 2006b Seasonality indices for regionalizing low flows. *Hydrol. Process.* **20**(18), 3851–3878.

Mahdavi, M. 2007 *Applied Hydrology*. University of Tehran Publication, Iran (in Persian).

Mamun, A. A., Hashim, A. & Daoud, J. I. 2010 Regionalisation of low flow frequency curves for the Peninsular Malaysia. *J. Hydrol.* **381**(1–2), 174–180.

MathWave Technologies. 2010 EasyFit 5.5, MathWave Technologies. Available at: http://www.mathwave.com

McCuen, R. H. 1998 *Hydrologic Analysis and Design* (2nd ed.). Prentice Hall, Upper Saddle River, NJ.

Modarres, R. 2008 Regional frequency distribution type of low flow in North of Iran by L-moments. *Water Resour. Manage.* **22**(7), 823–841.

Modarres, R. & Sarhadi, A. 2010 Frequency distribution of extreme hydrologic drought of southeastern semiarid region, Iran. *J. Hydrol. Eng.* **15**, 255–264.

Nathan, R. & McMahon, T. 1990 Identification of homogeneous regions for the purposes of regionalisation. *J. Hydrol.* **121**(1–4), 217–238.

NDMC 2005 What is drought? Understanding and defining drought: hydrological drought. University of Nebraska–Lincoln. Available at: http://drought.unl.edu/whatis/concept.htm (accessed 25 May 2011).

Nosrati, K. 2012 Regional analysis of hydrological drought in Sefidrood Drainage Basin using base flow index. *Journal of Range and Watershed Management, Iranian Journal of Natural Resources* **65**(2), 257–267 (in Persian).

Nosrati, K. & Shahbazi, A. 2008 Low flow estimation using hybrid method in Northeastern of Iran. *Iranian Journal of Natural Resources* **60**(3), 829–841.

Nosrati, K., Zehtabian, G., Mohseni Saravi, M. & Mahdavi, M. 2002 Best fitting distribution function for low flows. *Desert* **7**(1), 73–83 (in Persian).

Nosrati, K., Mohseni Saravi, M., Eslamian, S., Sharaifi, F. & Mahdavi, M. 2004 Identification of homogenous region for low flow frequency analysis. *Iranian Journal of Natural Resources* **57**(1), 45–57 (in Persian).

Rifai, H. S., Brock, S. M., Ensor, K. B. & Bedient, P. B. 2000 Determination of low-flow characteristics for Texas streams. *J. Water Resour. Plann. Manage.* **126**, DOI:10.1061/(ASCE)0733-9496(2000)126:5(310).

Rogerson, P. A. 2001 *Statistical Methods for Geography*. Sage, London, UK.

Ryu, J. H., Lee, J. H., Jeong, S., Park, S. K. & Han, K. 2011 The impacts of climate change on local hydrology and low flow frequency in the Geum River Basin, Korea. *Hydrol. Process.* **25**(22), 3437–3447.

Smakhtin, V. 2001 Low flow hydrology: a review. *J. Hydrol.* **240**(3–4), 147–186.

Stahl, K. 2001 *Hydrological drought - a study across Europe*. Universitätsbibliothek, Freiburg, Germany.

Stahl, K. & Demuth, S. 1999 Linking streamflow drought to the occurrence of atmospheric circulation patterns. *Hydrolog. Sci. J.* **44**(3), 467–482.

StatSoft 2008 STATISTICA: [data analysis software system], Version 8.0 for Windows update. StatSoft, Inc., Tulsa, OK.

Tabrizi, A. A., Khalili, D., Kamgar-Haghighi, A. A. & Zand-Parsa, S. 2010 Utilization of time-based meteorological droughts to investigate occurrence of streamflow droughts. *Water Resour. Manage.* **24**(15), 4287–4306.

Tallaksen, L. M. & van Lanen, H. A. J. 2004 *Hydrological Drought—Processes and Estimation Methods for Streamflow and Groundwater*. Elsevier, Amsterdam.

Tallaksen, L. M., Madsen, H. & Clausen, B. 1997 On the definition and modelling of streamflow drought duration and deficit volume. *Hydrolog. Sci. J.* **42**(1), 15–33.

Tasker, G. D. 1987 A comparison of methods for estimating low flow characteristics of streams. *JAWRA J. Am. Water Resour. Ass.* **23**(6), 1077–1083.

Vogel, R. M. & Kroll, C. N. 1989 Low-flow frequency analysis using probability-plot correlation coefficients. *J. Water Resour. Plann. Manage.* **115**(3), 338–357.

Vogel, R. M. & Kroll, C. N. 1990 Generalised low-flow frequency relationships for ungauged sites in Massachusetts. *JAWRA J. Am. Water Resour. Ass.* **26**(2), 241–253.

Vogel, R. M. & Kroll, C. N. 1992 Regional geohydrologic-geomorphic relationships for the estimation of low-flow statistics. *Water Resour. Res.* **28**(9), 2451–2458.

Yu, P. S., Yang, T. C. & Liu, C. W. 2002 A regional model of low flow for southern Taiwan. *Hydrol. Process.* **16**(10), 2017–2034.
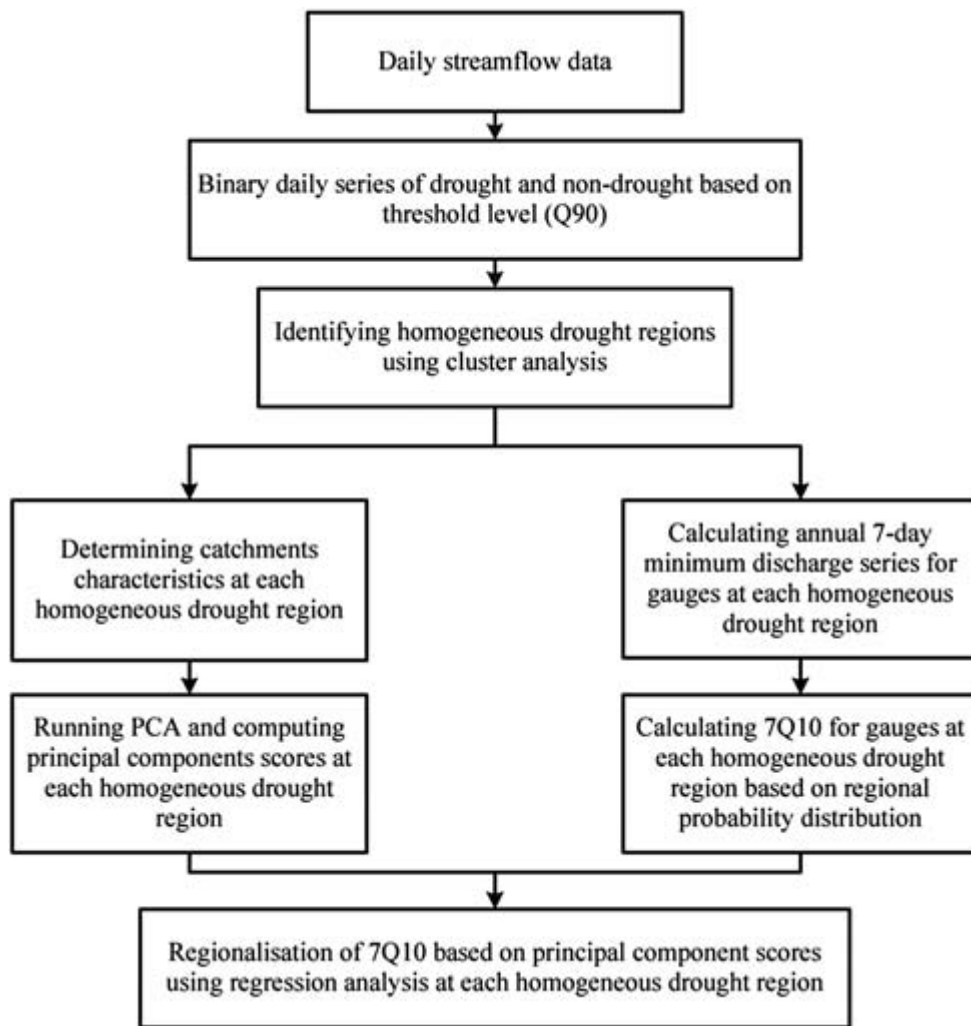
Figure 1 | The integrated approach for establishing the regional low flow analysis
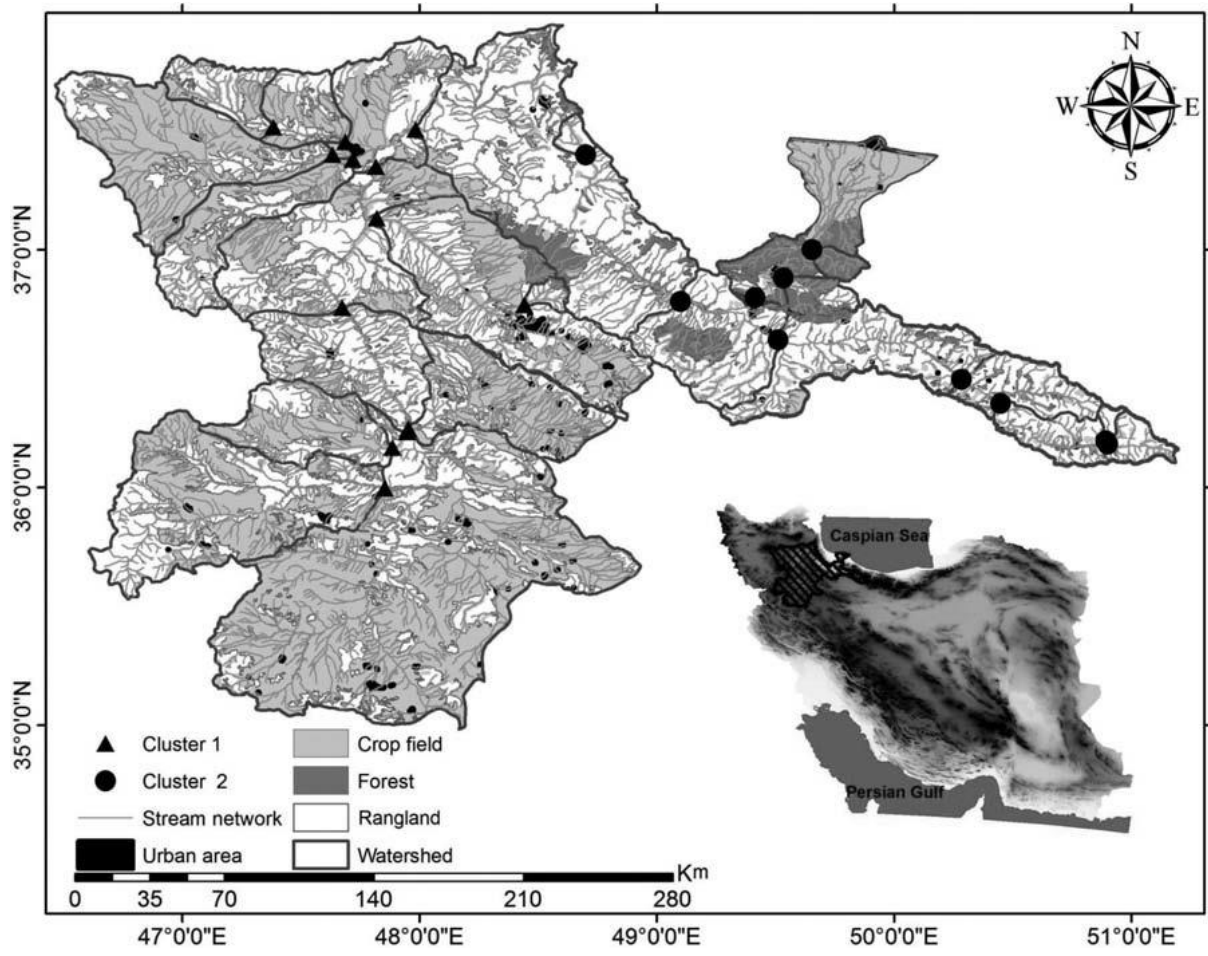
Figure 2 | Location map of the Sefidrood Drainage Basin and streamflow gauging stations in homogenous drought regions.
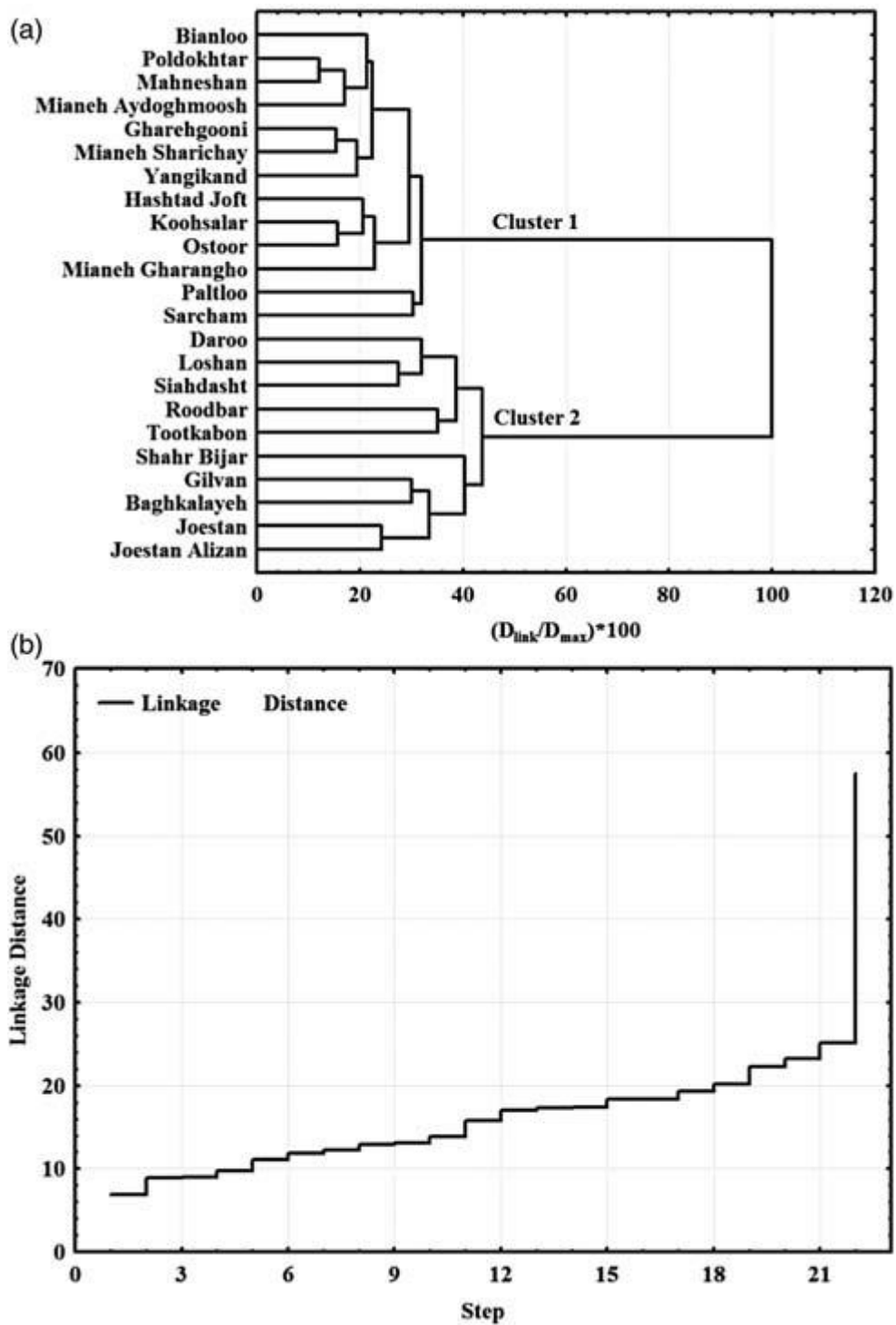
Figure 3 | Dendrogram showing clustering of streamflow gauging stations according to (a) binary drought series and (b) graph of amalgamation schedule.
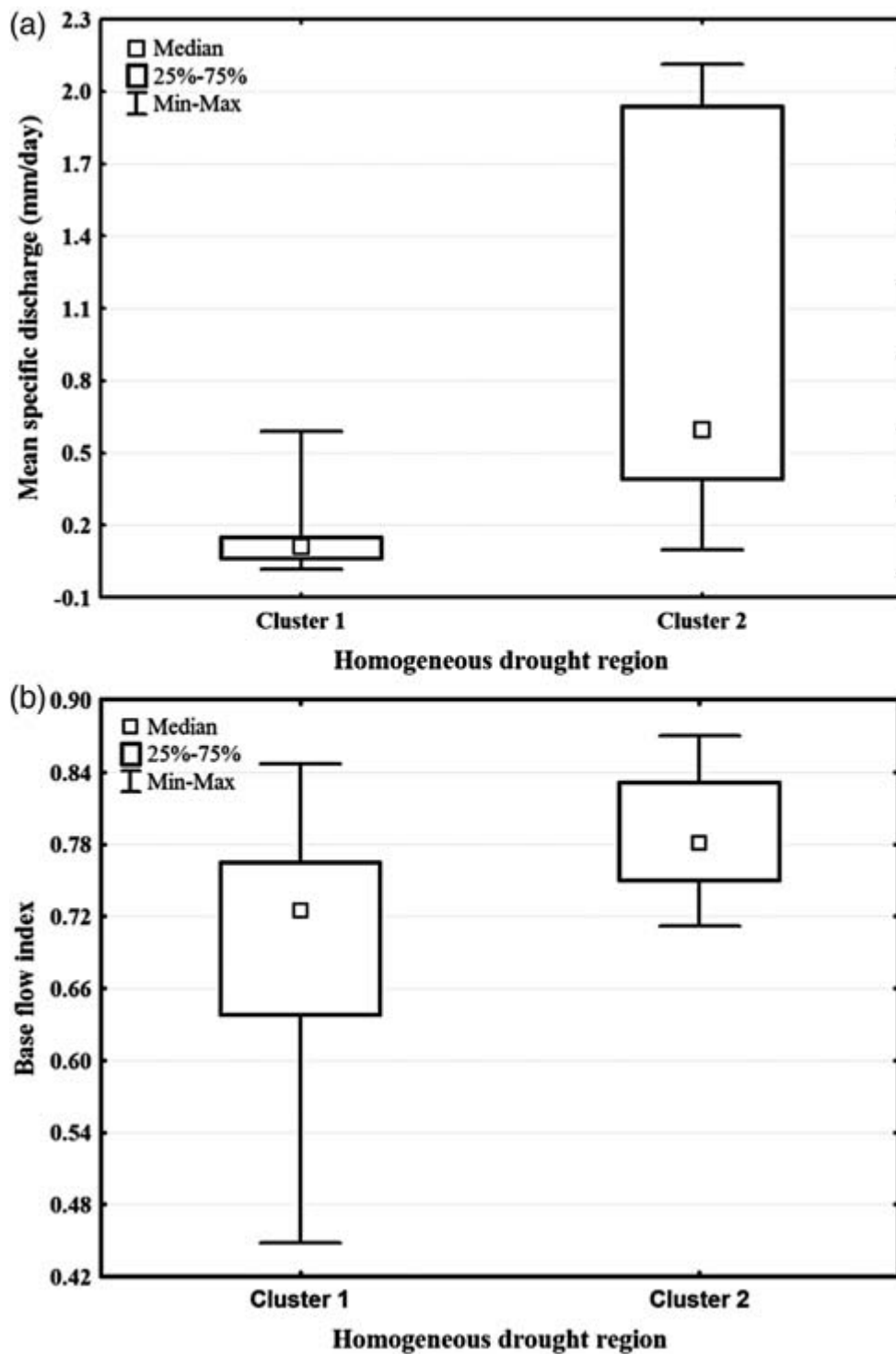
Figure 4 | Box plots of mean specific discharge and base flow index for all stations within the two homogenous drought regions.
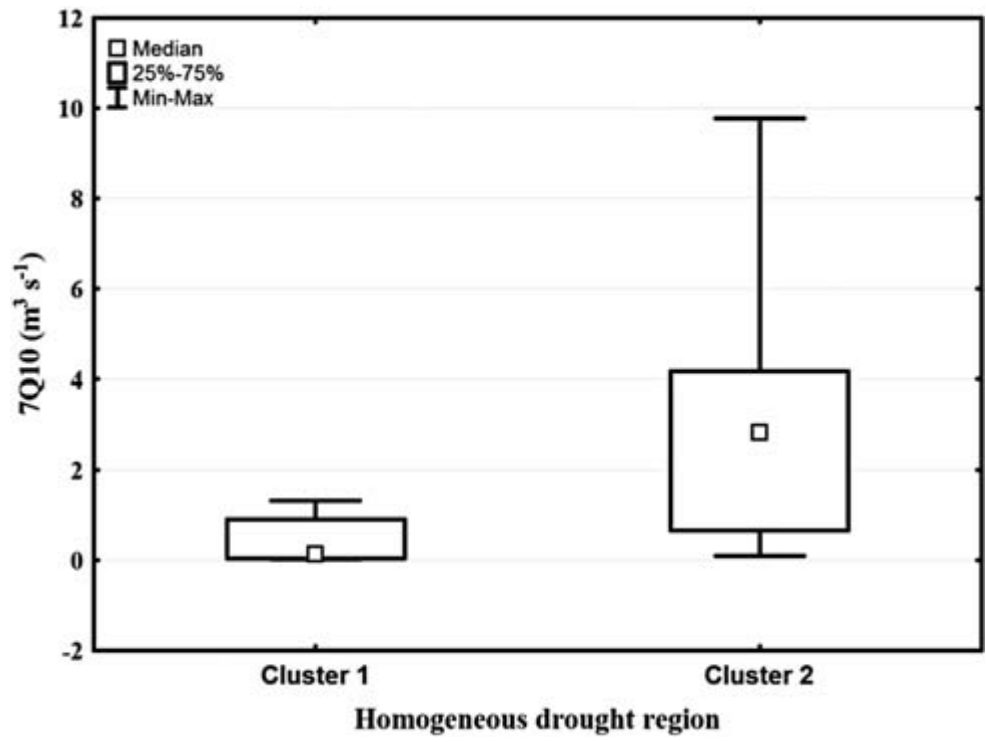
Figure 5 | Box plots of 7Q10 for all stations within the two homogenous drought regions.
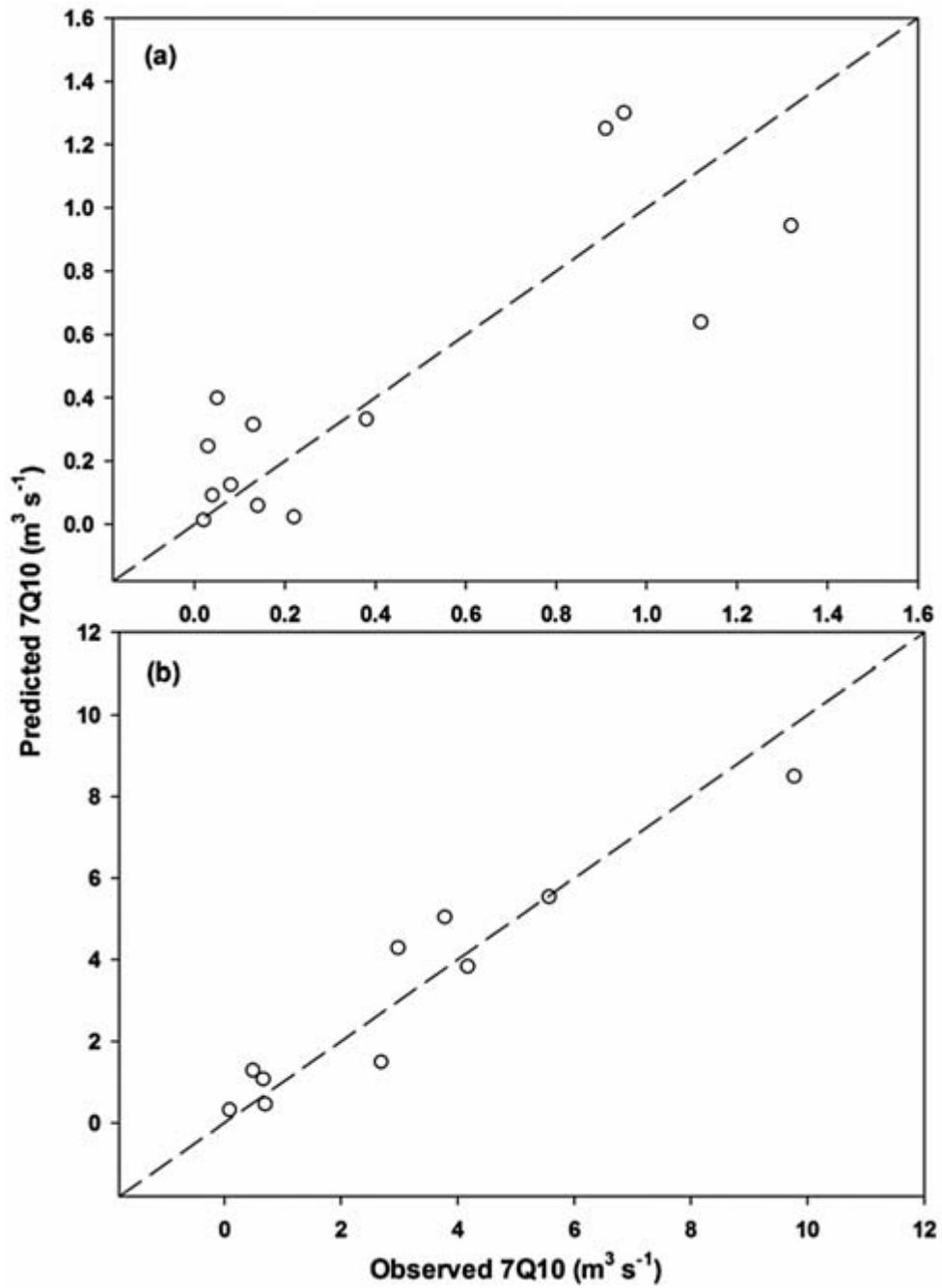
Figure 6 | Scatterplot of observed vs. predicted data points for low flow (7Q10): (a) homogenous drought cluster 1 and (b) homogenous drought cluster 2.

Table 1 | List of catchment characteristics used in this study

| Catchment characteristic	Source data /method	Unit	Description
Drainage area	30 m digital elevation model (DEM)	km^2	–
Drainage perimeter	30 m digital elevation model (DEM)	km	–
Mean slope gradient	30 m digital elevation model (DEM)	%	–
Mean, maximum and minimum elevation	30 m digital elevation model (DEM)	m	–
Main stream length	1:25,000-scale topographic map	km	The length of the longest channel present in the catchment
Summation of stream	1:25,000-scale	km	Summation of all streams lengths

lengths	topographic map		
Drainage density	1:25,000-scale topographic map	km km^{-2}	The ratio of the total length of all streams and area of the catchment
Length and width of rectangle-equivalent	Based on Roche definition (Mahdavi 2007)	km	According to Roche a catchment can be assumed to represent a rectangular shape with same area, which longitudinally coincides with that of the principal river axes called rectangle-equivalent
Circularity ratio	$F_c = 0.282M\ A^{-0.5}$ where M and A are the perimeter (km) and area (km^2) of the catchment, respectively	–	The shape factor of the catchments was described by circularity ratio
Time of concentration	Calculated by Kirpich equation: $t_c = 0.949(L^3/H)^{0.385}$ in which L is the length of the channel (km) and H is the difference in elevation between the points defining the upper and lower ends of the channel (m)	hr	The time required for runoff to travel from the hydraulically most distant point in the watershed to the outlet (McCuen 1998)
Mean annual precipitation	Isohyetal method (McCuen 1998)	mm	The area within each pair of adjacent isohyets used to weight the average annual precipitation associated with the adjacent isohyets
Urban and forest fraction	1:40,000-scale aerial photographs and 1:25,000-scale topographic map	%	

Geological formations with high, medium, low and very low infiltration capacity	1:250,000 scale geological map	%	Geological formations were classified based on hydraulic conductivity and mean annual specific discharge

Table 2 | Sum of computed ranks for the selected distributions in the two homogenous clusters and overall study area

Homogenous region	Probability distribution function									
	Gamma	GEV	GLOG	GPAR	LPIII	LN2	LN3	Normal	PIII	W3
Cluster 1	68	55	63	69	**40**	72	85	89	90	84
Cluster 2	66	56	51	51	63	**30**	72	37	73	51
Overall study area	134	111	114	120	**103**	**102**	157	126	163	135

LPIII, log-Pearson type-III; GEV, generalized extreme value; GLOG, generalized logistic; GPAR, generalized Pareto; PIII, Pearson type-III; LN2, 2-parameter log-normal; LN3, 3-parameter log-normal; W3, 3-parameter Weibull

Table 3 | Proportion of variance using varimax rotation and communality estimates of catchment characteristics for two homogenous drought regions

Variables	PC1	PC2	PC3	PC4	Communality estimates
Cluster 1 (four significant principal components)					
Drainage area (km^2)	**0.97**	0.01	0.19	0.11	0.98
Drainage perimeter (km)	**0.97**	0.09	0.14	0.16	1.00
Circularity ratio	**0.75**	0.26	−0.15	0.20	0.70
Length of rectangle-equivalent (km)	**0.95**	0.04	0.14	0.14	0.95
Width of rectangle-equivalent (km)	**0.83**	0.07	0.26	0.11	0.78
Maximum elevation (m)	0.47	0.36	**0.75**	0.22	0.95
Minimum elevation (m)	−0.44	**−0.81**	−0.15	−0.03	0.87

Mean elevation (m)	0.23	−0.29	**0.85**	0.24	0.92
Mean slope (%)	*0.58*	0.41	0.40	0.38	0.81
Summation of stream lengths (km)	**0.97**	0.02	0.16	0.12	0.97
Main stream length (km)	**0.95**	0.16	0.18	0.13	0.98
Drainage density (km km^{-2})	−0.15	−0.36	*−0.71*	0.34	0.77
Time of concentration (hr)	**0.96**	0.12	0.03	−0.20	0.97
Urban fraction (%)	0.27	0.35	*−0.67*	0.31	0.74
Very low ICGF (%)	−0.27	0.00	0.00	**−0.91**	0.91
Low ICGF (%)	0.23	**−0.85**	0.14	0.33	0.91
Medium ICGF (%)	−0.38	0.03	−0.26	0.29	0.29
High ICGF (%)	0.12	**0.88**	−0.05	0.21	0.83
Mean annual precipitation (mm)	**−0.76**	−0.04	0.33	−0.38	0.83
Eigenvalue	9.6	2.7	2.5	1.4	
% Total variance	50.5	14.3	13.1	7.2	
Cumulative % variance	50.5	64.8	77.8	85.0	

Cluster 2 (four significant principal components)

Drainage area (km^2)	**0.97**	0.02	−0.03	0.09	0.95
Drainage perimeter (km)	**0.95**	0.02	0.04	0.23	0.96
Circularity ratio	**0.72**	−0.08	0.25	0.08	0.92
Length of rectangle-equivalent (km)	**0.95**	0.02	0.04	0.24	0.96
Width of rectangle-equivalent (km)	**0.99**	0.06	−0.05	0.04	0.99
Maximum elevation (m)	0.26	−0.37	**0.82**	0.28	0.94
Minimum elevation (m)	−0.45	*−0.65*	0.39	−0.40	0.95
Mean elevation (m)	-0.12	*−0.65*	0.03	−0.09	0.98
Mean slope (%)	0.17	**0.91**	0.11	0.26	0.95
Summation of stream lengths (km)	**0.97**	0.02	−0.03	0.09	0.95
Main stream length (km)	**0.98**	0.02	0.01	0.18	0.99
Drainage density (km km^{-2})	−0.02	**0.81**	0.09	0.08	0.67
Time of concentration (hr)	**0.98**	0.01	−0.09	0.16	0.99

Urban fraction (%)	*0.58*	0.14	0.28	−0.05	0.43
Forest fraction (%)	−0.32	**0.83**	−0.41	−0.10	0.96
Very low ICGF[a] (%)	0.01	−0.20	**−0.87**	−0.05	0.81
Low ICGF (%)	0.08	0.33	0.16	**0.89**	0.93
Medium ICGF (%)	*0.70*	−0.17	−0.32	−0.24	0.68
High ICGF (%)	−0.40	−0.07	0.41	*−0.64*	0.95
Mean annual precipitation (mm)	**−0.75**	0.33	−0.01	−0.29	0.93
Eigenvalue	9.7	4.4	2.5	1.3	
% Total variance	48.4	21.9	12.6	6.6	
Cumulative % variance	48.4	70.4	82.9	89.5	

ICGF, infiltration capacity of geological formation

Bold and italic values indicate strong (>0.75) and moderate (0.75–0.50) loadings, respectively

Table 4 | Best set of prediction variables for low flow (7Q10) defined by stepwise multiple regression analysis of the two homogenous regions

| Homogenous region | Regression statistics | | | | | Regression parameters | | | | | |
	R^2	Adjusted R^2	F	p-value	SEE	Dependent variable	Predictor	B	t-value	Partial Correlation	VIF
Cluster 1	0.69	0.66	23	<0.001	0.27	7Q10	Intercept	0.40	3.9		
							PC1	0.39	3.4	0.83	1.0
Cluster 2	0.89	0.84	17	<0.001	1.21	7Q10	Intercept	3.09	8.1		
							PC1	2.10	5.1	0.91	1.0
							PC4	1.44	3.8	0.82	1.0
							PC3	1.19	3.0	0.77	1.0

SEE, standard error of estimate; B, raw regression coefficient; VIF, variance inflation factor