Regionalization of Drought Characteristics Using an Entropy Approach

Deepthi Rajsekhar[1]; Ashok K. Mishra[2]; and Vijay P. Singh, F.ASCE[3]

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

Abstract: Assessment and understanding of past climate is an important step for drought mitigation and water resources planning. In this study, streamflow simulated using the variable infiltration capacity (VIC) model was used for drought characterization for a time span of 1950–2000, and subsequently, regionalization was done for the state of Texas based on the annual drought severity level and drought duration. Droughts are regional in nature, and hence, identification of homogenous drought regions is important for investigating the drought characteristics within each of these regions. The concept of entropy was used for identification of homogenous regions based on drought severity and duration. A standardized version of mutual information, known as directional information transfer, was used for station grouping. The homogeneity of regions obtained was checked using L-moments. A total of eight regions were formed based on drought severity, and nine based on drought duration. Regions in west Texas were found to be critical in terms of severity, whereas east Texas showed the least severity. The longest drought duration was experienced in south Texas and lower valley zones, whereas the least drought duration was experienced in east Texas and the upper coast. Severely dry and extremely dry droughts were found to be restricted to the western and central parts of Texas. **DOI: 10.1061/(ASCE)HE.1943-5584.0000683.** © *2013 American Society of Civil Engineers.*

CE Database subject headings: Droughts; Entropy methods; Texas.

Author keywords: Drought regionalization; Entropy; Directional information transfer.

Introduction

Drought is an extended period of time during which a deficiency in precipitation is experienced. In many parts of the world, it is a normal, recurring feature of the climate and is therefore inevitable. It is a gradual phenomenon, and often it is difficult to identify the beginning or end of a drought (Wilhite and Glantz 1985). A drought can extend for just a few months, or it may persist for several years.

There is no universally accepted definition for droughts. They can be classified into meteorological, hydrological, groundwater, agricultural, and socioeconomic droughts (Mishra and Singh 2010). Droughts are the costliest of all the natural hazards, and hence have a huge impact on society, causing an average of $6–8 billion in global damages annually (Wilhite 2000). Adequate monitoring and planning is required for their effective mitigation.

Texas has been a consistently drought prone state. The number of drought years in each of the 10 geographic areas of Texas during the twentieth century was as follows: Trans-Pecos, 16 years; lower Rio Grande valley, 17; Edwards Plateau, 17; south central, 15; southern, 15; north central, 12; upper coast, 13; east Texas, 10; High Plains, 10; and Low Rolling Plains, 8 (Dunn 2011). There has been at least one serious drought in one part of the state or the other every decade of the twentieth century. This trend is likely to increase in the coming years because of the effect of global warming and climate change. Taking into account the importance of water management under conditions of extreme climate, this study focuses on hydrological droughts, wherein a deficit in streamflow will be the indicator of a drought event.

The regional nature of drought has been investigated by Sen (1980), Clausen and Pearson (1995), Hisdal and Tallaksen (2003), Byzedi and Saghafian (2009), Mishra and Singh (2009), and Mirakbari et al. (2010). The first step for a regional univariate or multivariate drought analysis is the identification of homogenous regions. A homogenous region can be defined as a group of stations with similar probability distribution functions of drought (Mirakbari et al. 2010). Similar water management schemes and drought planning can be developed for homogenous regions.

The common concept used in regional analysis of droughts is to classify weather stations that exhibit similarities in a statistical sense. There are several methods for performing regionalization. Some of the common approaches for regionalization in hydrology include the method of residuals (MOR) approach (Choquette 1988), the region of influence (ROI) approach (Zrinji and Burn 1994, 1996), the principal component analysis (PCA) approach (Singh and Singh 1996), and cluster analysis and its extensions (Rao and Srinivas 2006a, b; Isik and Singh 2008; Srinivas et al. 2008; Satyanarayana and Srinivas 2011). The MOR approach delineates regions in an arbitrary fashion, and regions are arranged to match existing political, geographic, or hydrologic boundaries (Rao and Srinivas 2006b). The ROI approach defines groups of sites in a flexible manner such that each station has its own region. Although the method overcomes the inconsistencies that may occur on the boundaries of groups (Acreman and Wiltshire 1989), there are no strict mathematical solutions for the selection and weighing of variables (Bobee and Rasmussen 1995). The PCA approach determines the net effect of each variable on the total variance of the

[1]Graduate Student, Biological and Agricultural Engineering (BAEN) Dept., Texas A&M Univ., College Station, TX 77840 (corresponding author). E-mail: deepthir86@gmail.com

[2]Assistant Research Scientist, BAEN Dept., Texas A&M Univ., College Station, TX 77840.

[3]Caroline and William N. Lehrer Distinguished Chair in Water Engineering and Professor, Civil and Environmental Engineering Dept. and BAEN Dept., Texas A&M Univ., College Station, TX 77840.

Note. This manuscript was submitted on October 12, 2011; approved on July 5, 2012; published online on August 6, 2012. Discussion period open until December 1, 2013; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Hydrologic Engineering***, Vol. 18, No. 7, July 1, 2013. © ASCE, ISSN 1084-0699/2013/7-870-887/$25.00.

data set, and then tries to explain the maximum possible variance using the minimum number of variables. The PCA approach has a disadvantage in that there is no criterion against which to check the results. The groups formed are highly subjective in nature. Although no single procedure has been identified as the most acceptable one, the use of various clustering algorithms seems to be popular. The hierarchical clustering method (Nathan and McMahon 1990; Burn et al. 1997) proceeds by either agglomeration or division of existing clusters. The partitioning clustering method (Bhaskar and O'Connor 1989; Burn and Goel 2000) determines all the clusters at one go. Rao and Srinivas (2006b) noted that the use of simple clustering methods might not yield regions that satisfy all three heterogeneity measures, H_1, H_2, and H_3, of Hosking and Wallis (1997). However, the use of hybrid clustering techniques (Rao and Srinivas 2006b; Srinivas et al. 2008) gave considerably better results. Apart from the aforementioned methods, others like kriging (Chokmani and Ouarda 2004), self-organizing feature maps (Jingyi and Hall 2004), and a combination of clustering algorithms with flow duration curves (Isik and Singh 2008) have also been employed.

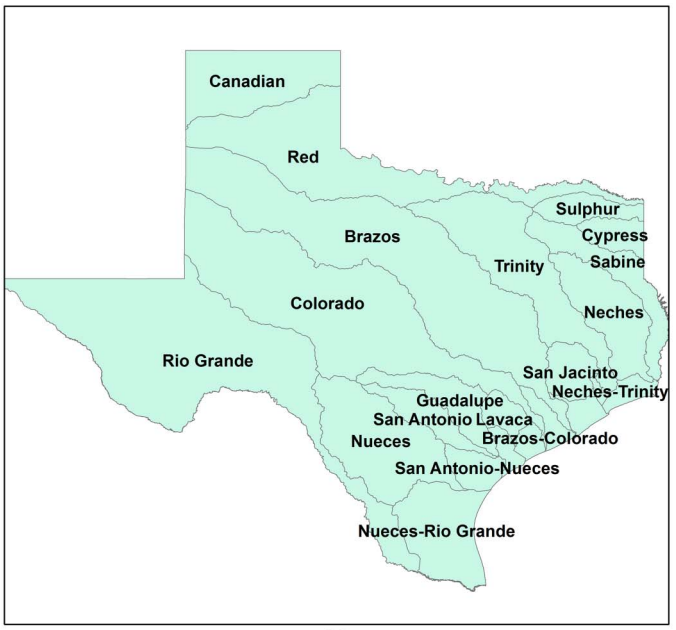
Selection of a suitable similarity measure is important in clustering. Mostly, clustering techniques use a simple linear measure like Pearson correlation as a similarity measure for grouping. In this study, the possibility of using a mutual information-based index known as directional information transfer (DIT) for identification of homogenous regions was explored. This measure is not only sensitive to nonlinear dependencies, but it is also unique because of its information theory background (Kraskov and Grassberger 2009). It has a threefold advantage over other dependence measures in that it gives an idea about (1) information content at a station, (2) the amount of information transferred between stations and the amount lost, and (3) relationships among stations based on information transmission characteristics (Yang and Burn 1994).

The study basically focuses on understanding the spatial distribution of drought characteristics. Further, the identification of homogeneous drought regions will be required for conducting regional drought frequency analyses. An areal zoning of the study region based on various drought properties was conducted using a methodology based on entropy theory using Texas as the case study area. The methodology is based on an index developed by Yang and Burn (1994) for design of a data collection network. The same principle has been extended for the grouping of stations. The application of this method in the context of regionalization has not been explored until now.

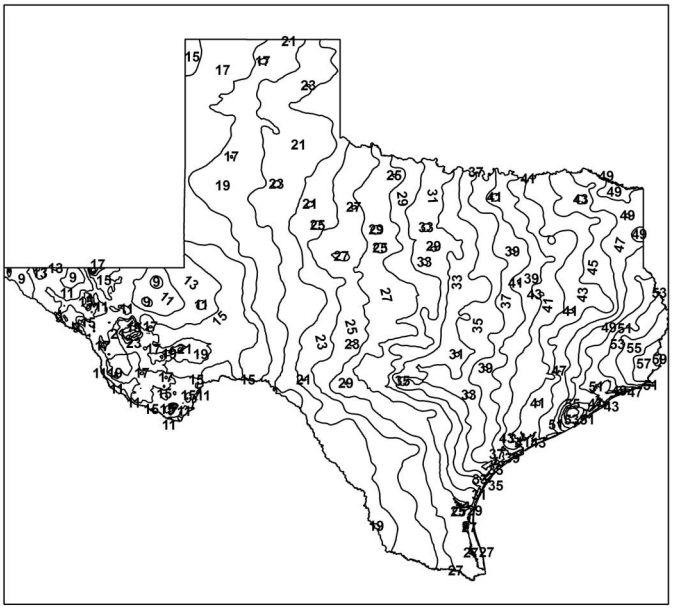
The objectives of the paper are therefore to (1) apply the variable infiltration capacity (VIC) model for streamflow drought analysis, (2) do regionalization of the annual drought severity levels and drought duration for the state of Texas, and (3) identify critical regions within Texas using entropy. Knowledge of the spatial variability of drought properties will help in developing a prototype water management scheme for each region separately.

Study Area

The area considered for this study is the state of Texas. There are five distinct climate zones in Texas showing the variation from arid to subtropic humid zones. The varied physiography in the state of Texas with forests in the east, coastal plains in the south to the elevated plateaus, and basins in the north and west, results in a wide variety of weather throughout the year (Benke and Cushing 2005). The land surface elevation follows a decreasing trend from west to east with an arid climate zone covering higher elevation areas, whereas most of the subtropic humid zone and parts of the subtropic semihumid zone covers the low lying regions in Texas. There are 13 major river basins in Texas that vary greatly in size, shape, and stream patterns. Climate, particularly rainfall and evaporation, strongly controls the flows of rivers and streams in Texas. The region is traversed by a strong decreasing rainfall gradient from east to west and a temperature gradient from north to south that strongly influences vegetation, land use, and river flow. In the Sabine River basin in east Texas, mean annual rainfall is nearly 152.4 cm (60 in.) and annual evaporation is less than 177.8 cm (70 in.), whereas in the Rio Grande basin in west Texas, mean annual rainfall ranges from 20.32–50.8 cm (8–20 in.) and annual evaporation is as much as 266.7 cm (105 in.). Therefore, east Texas rivers flow year-round, whereas most of the west Texas streams flow only part of the year (Bureau of Economic Geology 1996). Fig. 1(a) shows the river basin map of Texas and the precipitation (annual average in inches) gradient within the state. Fig. 1(b) shows the five major climate



(a)



(b)

Fig. 1. (a) River basin map of Texas; (b) precipitation pattern in Texas

J. Hydrol. Eng. 2013.18:870-887.

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

Table 1. Information on Validation Stations within Texas

Station name	Station identification	Latitude	Longitude	Validation period	Climate zone
Pecos River at Pecos	8420500	31.436	−103.467	1951–1952	Arid
Canadian River near Amarillo	7227500	35.471	−101.88	1981–1982	Continental
Prairie Dog Town fork Red River near Wayside	7297910	34.837	−101.414	1968–1969	Continental
Canadian River near Canadian	7228000	35.935	−100.371	1965–1966	Continental
Independence Creek near Sheffield	8447020	30.452	−101.733	1975–1976	Semiarid
Nueces River near Asherton	8193000	28.5	−99.682	1959–1960	Semiarid
Colorado River near Gail	8117995	32.628	−101.285	1989–1990	Subtropical Semihumid
Colorado River near Stacy	8136700	31.494	−99.574	1969–1970	Subtropical Semihumid
Millers Creek near Munday	8082700	33.329	−99.465	1972–1973	Subtropical Semihumid
Medina River near Macdona	8180700	29.335	−98.689	1982–1983	Subtropical Semihumid
Cowhouse Creek at Pidcoke	8101000	31.285	−97.885	1955–1956	Subtropical Semihumid
Perdido Creek at M 622 near Fannin	8177300	28.752	−97.317	1979–1980	Subtropical Semihumid
Los Olmos Creek near Falfurrias	8212400	27.2645	−98.136	1967–1968	Subtropical Semihumid
Lake Fork Creek near Quitman	8019000	32.763	−95.463	1982–1983	Subtropical humid
Kickapoo Creek near Onalaska	8066170	30.907	−95.088	1991–1992	Subtropical humid
Vince Bayou at Pasadena	8075730	29.6947	−95.216	1973–1974	Subtropical humid

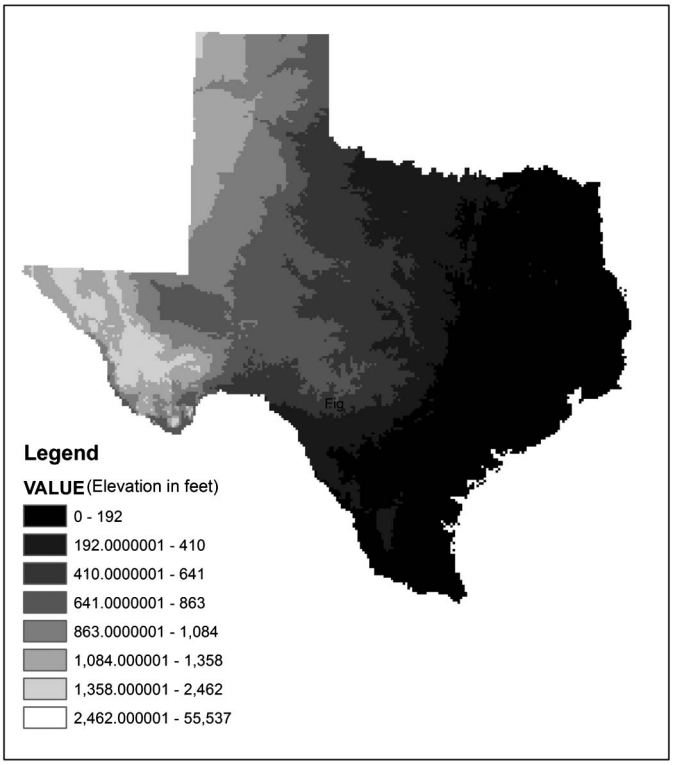


Fig. 2. Elevation map of Texas

zones within Texas, namely arid, semiarid, subtropical semihumid, subtropical humid, and continental steppe, and the locations of stream gauge stations used for validating the streamflow obtained from the VIC model. Table 1 gives details of the validation stations. Fig. 2 shows the elevation map of Texas. Fig. 3 shows the locations of the validation stations within various climate zones in Texas.

Model Description

Because finescale data is essential to account for spatial heterogeneity of droughts, it might not be wise to use stream gauge data, because stream gauges integrate over large spatial areas and thus do not account for the spatial variability of droughts (Andreadis et al. 2005). To avoid this problem and to overcome the unavailability of long-term continuous streamflow data all over Texas, a land surface model called the VIC model (Liang et al. 1996) was used to simulate streamflow for a period of 1950–2000, and results were validated against observed values from several USGS stream gauges. This particular model was chosen because it focuses on simulating hydrological processes relevant to the water and energy balance over the land surface for studying the effects of climate changes on streamflow generation. Distinguishing characteristics of the model include the subgrid variability in land surface vegetation classes, subgrid variability in the soil moisture storage capacity, and drainage from the lower soil moisture zone (base flow) as a nonlinear recession. The VIC model has been well calibrated and applied in a number of large river basins over the continental United States and the globe, and has participated in the World Climate Research Program (WCRP) Intercomparison of Land Surface Parameterization Schemes (PILPS) project and the North American Land Data Assimilation System (NLDAS), where it has performed well relative to other schemes and to available observations (Bowling et al. 2003a, b; Lohmann et al. 1998). The VIC-3L is a large-scale land surface model and is used for simulating land-atmosphere fluxes by solving water and energy balance at a daily or subdaily temporal scale (Liang et al. 1994). The land surface is essentially divided into grids of specified resolution. Each of these cells will be simulated independent of one another. Land surface is divided into different vegetation covers in such a way that multiple vegetation classes can exist within a cell. The soil moisture distribution, infiltration, drainage between soil layers, surface runoff, and subsurface runoff are all calculated for each land cover tile at each time step. Then for each grid cell, the total heat fluxes (latent heat, sensible heat, and ground heat), effective surface temperature, and the total surface and subsurface runoff are obtained by summing over all the land cover tiles weighted by fractional coverage. It should thus be noted that the VIC model does not account for the interflow between grids. Because of the absence of observed data for evaporation, soil moisture, and runoff for each grid, to evaluate the model simulation results, a routing model should be used as a postprocessing tool to produce streamflow at the points of interest (in this study, these points would be the selected USGS stations). Details of the routing model are given subsequently in the section. The vegetation parameters considered by the model include root depth, root fraction, leaf area index (LAI), stomatal resistance, and albedo. The main soil parameters include hydraulic conductivity, thickness of each soil layer, soil moisture diffusion parameters, initial soil moisture, bulk density, and particle density.

The earlier version of VIC, which was named VIC-2L, had only two soil layers, and Liang et al. (1996) found out that the model

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

tended to underestimate the evaporation because of the low soil moisture in its upper soil layer. The main cause of this error was the lack of a mechanism for moving moisture from the lower to the upper soil layer. The VIC-2L version was then modified to allow diffusion of moisture between soil layers, and to have an additional 10-cm-thin soil layer on the top of the previous upper soil layer. In this way, the three-layer VIC model (VIC-3L) was generated, and the VIC-3L framework has been used ever since. Thus, typically in the VIC-3L model, the soil is partitioned into three layers vertically with variable soil depths. The VIC model has been widely used, particularly for streamflow and soil moisture simulations. Abdulla et al. (1996), Nijssen et al. (1997, 2001), and Lohmann et al. (1998) used VIC primarily for streamflow simulation. Sheffield et al. (2004), Andreadis and Lettenmaier (2006), Sheffield and Wood (2008), and Shukla and Wood (2008) demonstrated the use of VIC-simulated soil moisture and runoff in the context of droughts.

Because the grid-based VIC model simulates the time series of runoff only for each grid cell, which is nonuniformly distributed within the cell, a stand-alone routing model (Lohmann et al. 1996, 1998) was employed to transport grid cell surface runoff and base flow to the outlet of that grid cell and then into the river system. In this routing scheme, the surface runoff simulated by VIC in each grid cell was transported to the outlet of the grid cell using a unit hydrograph approach. Then, runoff from each grid cell was routed through the channel using a linearized Saint-Venant equation.

In this study, the model was run separately for each of the 23 river basins in Texas, and once the streamflow simulations within each grid cell were obtained, the routing model was employed to transport grid cell surface runoff and base flow to the outlet of that grid cell and then into the river system. In the routing model, water is never allowed to flow from the channel back into the grid cell. Once it reaches the channel, it is no longer part of the water budget. A linear transfer function model characterized by its internal impulse response function was used to calculate the within-cell routing. Then by assuming all runoff exits a cell in a single flow direction, a channel routing based on the linearized Saint-Venant equation was used to simulate discharge at the basin outlet.

Data

For this study, the VIC model for streamflow simulation was run at $1/8$ degree resolution, and hence all input files, including forcing files, soil, and vegetation parameters have this resolution. This is the default resolution at which the VIC model runs (Salathe 2003). This resolution was chosen by also taking into consideration the availability of gridded daily forcing data of precipitation (mm), maximum and minimum temperature (°C), and wind speed (m/s) that is needed to drive the model at $1/8$ resolution from Maurer et al. (2002), who provided a database for 15 delineated basins in the United States, Canada, and Mexico. The time period of data used was the latter half of the twentieth century: 1949–2000. The year 1949–1950 was considered the spin up year for the model. Apart from forcing data, soil and land cover data are also required by the VIC model. The soil characteristics that will not be considered for calibration were taken from gridded $1/8$ degree data sets developed as part of the LDAS project (Mitchell et al. 1999). Within the conterminous United States, these data sets are based on the 1-km-resolution data set produced by Pennsylvania State University (Miller and White 1998). Soil texture in the LDAS data set is divided into 16 classes for each of the 11 layers, inferring specific soil characteristics (e.g., field capacity, wilting point, and saturated hydraulic conductivity) based on the work of Cosby et al. (1984), Rawls et al. (1993), and Reynolds et al. (2000). These LDAS data sets were used to specify the relevant soil parameters required by the VIC model directly. For the remaining soil characteristics (e.g., soil quartz content), values were specified using the soil textures from the 1-km database, which were then indexed to published parameter values [the primary source was Rawls et al. (1993)], and aggregated to the $1/8$ degree model resolution. Vegetation parameters needed were also obtained from LDAS. Land cover characterization was based on the University of Maryland global vegetation classifications described by Hansen et al. (2000), which has a spatial resolution of 1 km, and a total of 14 different land cover classes. From these global data, the land cover types present in each $1/8$ grid cell in the model domain and the proportion of the grid cell occupied by each as described by Maurer et al. (2000) were identified. The LAI needed was derived from the gridded ($1/4$ degree) monthly global LAI database of Myneni et al. (1997), which was inverted using the Hansen et al. (2000) land cover classification to derive monthly mean LAIs for each vegetation class for each grid cell.

The data needed for the routing scheme include a fraction file, flow direction file, Xmask file, flow velocity and diffusion files, and unit hydrograph file. ArcMap was used for the preparation of files, and the digital elevation model (DEM) files needed for creating the required files were obtained from the USGS hydro 1k data sets.

Methodology

Drought Classification Using Standardized Streamflow Index

Before using streamflow data for drought classification and further analysis, the time series was checked for stationarity. The assumption of stationarity might no longer be valid for time series of hydrological variables because of substantial climate change brought about by human intervention (Milly et al. 2008). Stationarity is the property of time series by virtue of which mean, variance, skewness, spectrum, and probability distribution of the time series do not vary with time. If the time series is found to be nonstationary, it should be rendered stationary by means of some transformation such as data differencing or detrending, which is given by

$$X_i^* = X_{i+1} - X_i, \qquad i = 1, 2, \ldots, N-1 \tag{1}$$

where X_i, $i = 1, 2, \ldots, N$ = monthly time series under consideration. To test for stationarity, two tests were used—the augmented Dickey Fuller (ADF) test proposed by Dickey and Fuller (1979), which tests for the difference stationarity, and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test proposed by Kwiatkowski et al. (1992), which tests for trend stationarity. The ADF tests were conducted through ordinary least squares (OLS) estimation of regression models with either an intercept or a linear trend. The KPSS test complemented the Dickey-Fuller unit root test (Kwiatkowski et al. 1992). In this test, the series was decomposed into the sum of a deterministic trend, a random walk, and a stationary error. The null hypothesis for the test was that the intercept was a fixed element.

Both the tests were carried out at the 5% significance level. If the time series failed to pass the KPSS test, detrending was carried out to remove the trend component from the series. If it failed the ADF test, data differencing was carried out to bring about stationarity. Once the stationarity tests were conducted on the monthly streamflow time series, it was used for drought classification using theory of runs.

J. Hydrol. Eng. 2013.18:870-887.

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

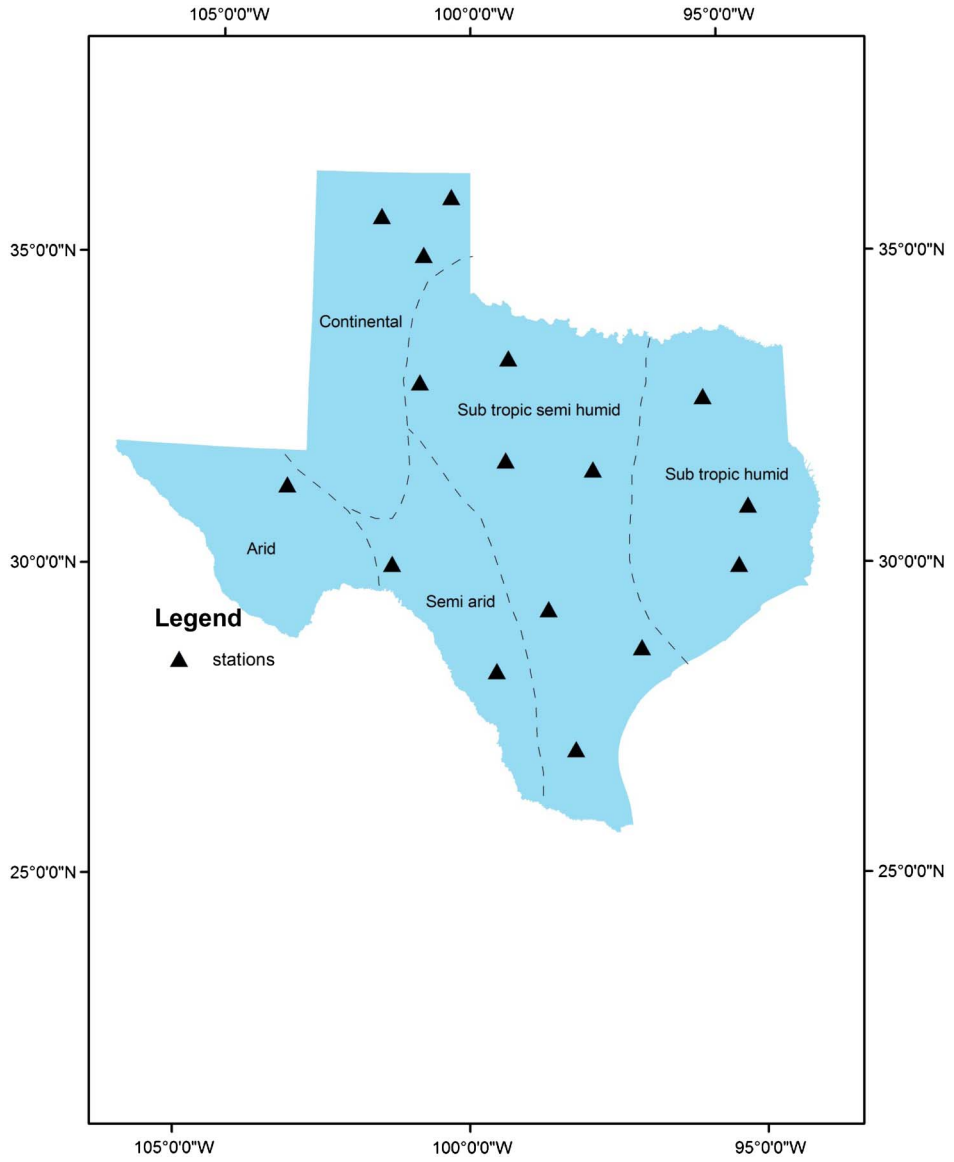


Fig. 3. Locations of validation stations within different climate zones in Texas

Fig. 4 describes drought characteristics for a drought event using the theory of runs. A drought event is characterized by severity, duration, and magnitude (Mishra and Singh 2010). For any drought event, the cumulative deficit of the variable of interest during the drought event is defined as drought severity. Drought duration is the time between the onset and the end of a drought event. Drought magnitude is the average deficit per unit duration. In this study, drought duration and severity were considered.

The theory of runs was used for deriving drought characteristics from the streamflow time series. This method has been widely used in the field of hydrology. Yevjevich et al. (1967), Rodriguez-Iturbe (1969), Saldarriaga and Yevjevich (1970), Millan and Yevjevich (1971), Guerrero-Salazar and Yevjevich (1975), and Sen (1976, 1977) were among the first few works that applied the run theory in hydrology. A run is defined as a portion of the time series of the drought variable X_t in which all values are either above or below a threshold level X_0. Accordingly, it can be called a positive or a negative run. The threshold level may be constant or may vary with time. Thus, the drought characteristics essentially depend on the threshold chosen (Mishra and Singh 2010). In this study, the drought variable chosen was the standardized streamflow index (SSFI). The concept of SSFI is statistically similar to that of the standardized precipitation index (SPI) introduced by McKee et al. (1993), and has been applied by Modarres (2007). Shukla and Wood (2008) used a standardized runoff index (SRI) as a complement to the SPI to assess hydrological aspects of a drought. Table 2 gives the classification of events based on the SSFI values (Modarres 2007). Following this classification, a threshold value of −0.99 was chosen, because any value below that indicates the onset of a dry event.

The calculation of SSFI involves the following steps: (1) a suitable probability distribution is fitted to the monthly streamflow time series for the time period 1950–2000; (2) from the fitted frequency distribution, the cumulative probability distribution of streamflow is obtained; and (3) cumulative probability is transformed to a standard normal variate of zero mean and unit standard deviation. This will be calculated from a numerical approximation to the normal cumulative distribution function (CDF). The approximation given by Abramowitz and Stegun (1964) was used to obtain

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

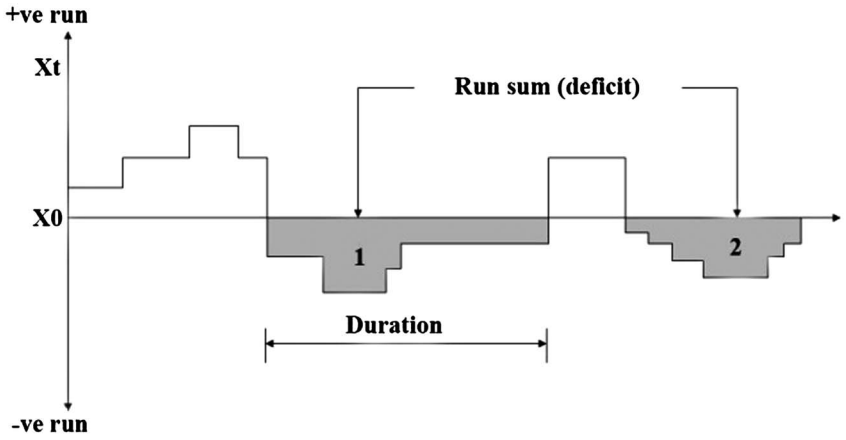


Fig. 4. Drought characteristics using theory of runs

Table 2. SSFI Classification

SSFI value	Classification
2.0 or more	Extremely wet
1.5 to 1.99	Very wet
1.0 to 1.49	Moderately wet
−0.99	Near normal
−1.0 to −1.49	Moderately dry
−1.5 to −1.99	Severely dry
−2.0	Extremely dry

the standard normal probability distribution function (PDF). The approximation for $\phi(x)$ for $x > 0$ is given by

$$\phi(x) = 1 - \varphi(x)(b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4 + b_5 t^5) + \varepsilon(x), \quad t = \frac{1}{1 + b_0 x} \tag{2}$$

where $\varphi(x)$ = standard normal PDF; $b_0 = 0.2316419$; $b_1 = 0.319381530$; $b_2 = -0.356563782$; $b_3 = 1.781477937$; $b_4 = -1.821255978$; and $b_5 = 1.330274429$.

This is the z-score, and conceptually it represents the number of standard deviations above or below that an event is from the mean (McKee et al. 1993). Thus, essentially SSFI for a given series is given as

$$\text{SSFI} = \frac{F_i - \bar{F}}{\sigma} \tag{3}$$

where F_i = flow rate in time interval i; $\bar{F}$ = mean of the series; and σ = standard deviation of the series.

In this study, for each of the five climatic regions, considering a number of previous studies such as Zaidman et al. (2001), Kroll and Vogel (2002), McMahon et al. (2007), Shukla and Wood (2008), and Nalbantis and Tsakiris (2009), the lognormal distribution was selected for fitting monthly streamflow data. The two-parameter lognormal distribution was found to fit well for all the stations considered. The quantile plots and Kolmogorov-Smirnov (K-S) test were considered for assessing the goodness of fit. Table 3 gives the results of the K-S test for the goodness of fit at the 5% significance level. Fig. 5 shows the quantile–quantile plot for two-parameter lognormal distribution used to fit streamflow at the selected stations.

Table 3. Values of the Kolmogorov-Smirnov Test at 5% Significance Level for Two-Parameter Lognormal Distribution at Selected Stations

Station	Climate zone	*P*-value	k-s stat
ID 08101000	Subtropical semihumid	0.0735	0.1280
ID 07227500	Continental	0.0684	0.2629
ID 08193000	Semiarid	0.0738	0.2546
ID 08420500	Arid	0.5500	0.1267
ID 08019000	Subtropical humid	0.4597	0.1689

Regionalization based on Directional Information Transfer

Regionalization is the process of identifying homogenous regions. In this context, a homogenous region comprises an area that has similar hydrologic response. This is generally done by grouping similar objects. Traditionally, a clustering algorithm is used for the purpose of dividing a set of feature vectors into groups, such that members within a cluster are as similar as possible, and members of different clusters are as dissimilar as possible (Rao and Srinivas 2006b). Thus, the most important aspect of any clustering algorithm is the selection of a similarity or dissimilarity measure. One of the most commonly used similarity measures is the Pearson correlation coefficient, which cannot be used as a nonlinear dependence measure.

In this study, an entropy-based index, known as DIT, was used for the grouping of grids into homogenous regions. This index is based on mutual information that measures information transfer among the stations. Entropy can be used to measure the information content of observations, and mutual information can be used to measure the information transfer. Thus, entropy and mutual information provide a threefold measure of information at a station, information transfer and loss of information between stations, and description of relationships among stations according to the information transfer between them (Yang and Burn 1994). This makes it unique from other conventional similarity measures. The following section discusses the basic concepts of entropy and directional information transfer.

Entropy Concepts

Entropy, first introduced in the field of information theory by Shannon (1948), is defined for a random variable X as (Lathi 1968)

$$H(X) = \sum_{i=1}^{k} P(x_i)\log_2 P(x_i) \tag{4}$$

J. Hydrol. Eng. 2013.18:870-887.

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

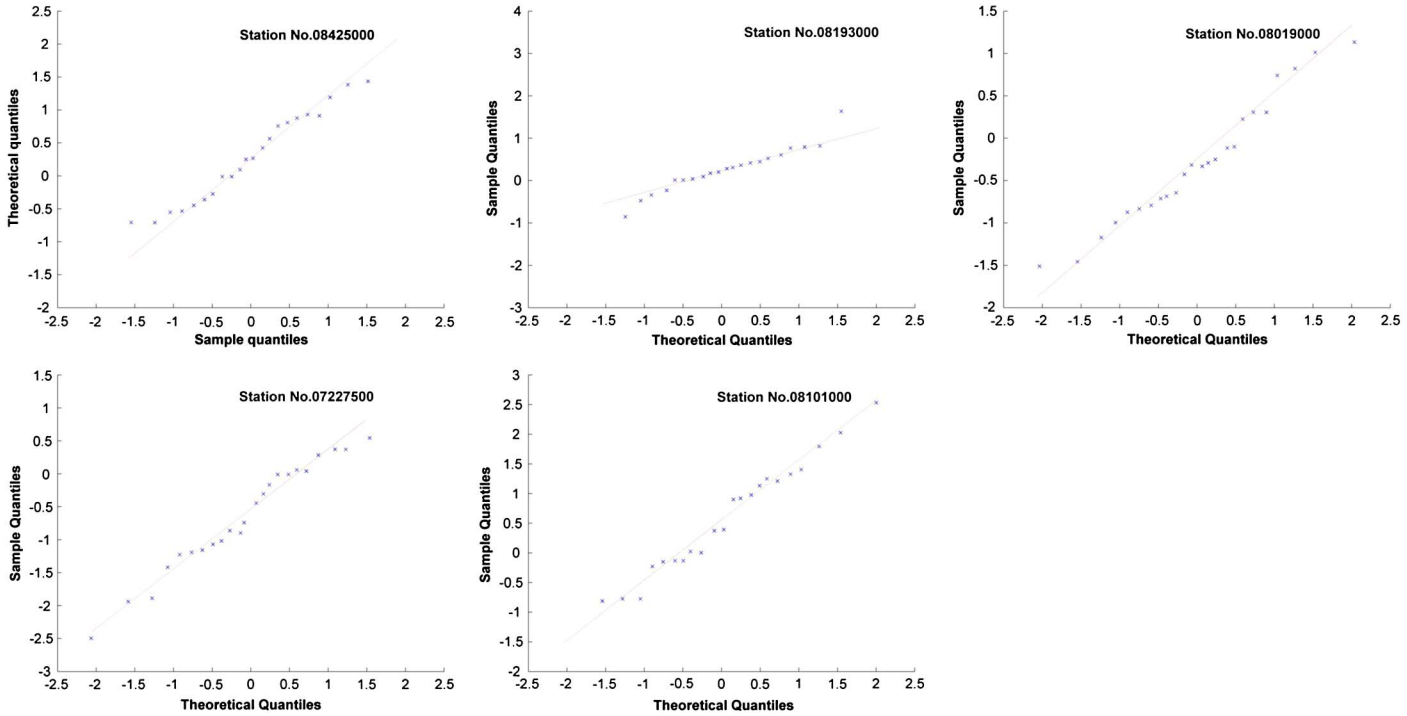


Fig. 5. Q-Q plots for two-parameter lognormal distribution used to fit streamflow at selected stations

where $P(x_i)$ = probabilities associated with the events $X = x_i$; and k = total number of class intervals or bins. The probabilities $P(x_i)$ can be calculated using histogram. The value of $H(X)$ is the marginal entropy of X, which means the measure of information contained in X. If X is a deterministic variable, then the probability that it will take on a certain value is one, and the probabilities of all other alternative values are zero. Hence, $H(X) = 0$, which will be the lower limit of the range of values $H(X)$ can take on. In contrast, when all $x_i s$ are equally likely, i.e., the variable is uniformly distributed ($P(x_i) = 1/N$, $i = 1, 2, \ldots, n$), $H(X) = \log_2 n$, which will be the upper limit for $H(X)$. If two random variables (X, Y) are considered, the mutual information or the measure of information transfer between them can be computed as (Lathi 1968)

$$T(X, Y) = H(X) - H(X|Y) \tag{5}$$

where $H(X|Y))$ = information lost during transmission, which can be estimated as

$$H(X|Y) = \sum_{i,j} P(X_i, Y_j) \log_2 \frac{P(X_i, Y_j)}{P(Y_j)} \tag{6}$$

where $P(X_i, Y_j)$ = joint probability distribution; $P(Y_j)$ = marginal distribution of random variable Y; and i and j = class intervals corresponding to X and Y, respectively. A simple histogram method can be used to estimate the required marginal probabilities. A bivariate histogram of the paired random variables X and Y and the associated contingency table can be used to estimate the joint probability required for the calculation of transinformation. Now suppose that there are n observations of events (X_i, Y_j), and n_{ij} denotes the number of times X_i occurred and Y_j was caused. In other words,

$$n_i = \sum_j n_{ij}; \qquad n_j = \sum_i n_{ij}; \qquad n = \sum_{i,j} n_{ij} \tag{7}$$

where n_i = number of times Xi occurred; n_j = number of times Yj was caused; and n = total number of observations. The contingency tables used for calculation of joint probabilities would have relative frequencies as the entries, calculated by

$$P(X_i) = \frac{n_i}{n}, \qquad P(Y_j) = \frac{n_j}{n}, \qquad P(X_i, Y_j) = \frac{n_{ij}}{n} \tag{8}$$

The bin size estimation for the histogram can be based on Sturges' formula (Sturges 1926), given by

$$k = \log_2 n + 1 \tag{9}$$

where n = sample size; and k = number of bins.

It should be noted that

$$\begin{aligned} T(Y, X) &= H(X) + H(Y|X) - H(X|Y) - H(Y|X) \\ &= H(X) - H(X|Y) = T(X, Y) \end{aligned} \tag{10}$$

Hence, it can be seen that the transinformation term is symmetric.

Mutual information has been used as a similarity measure for clustering purposes (Kraskov et al. 2005; Kraskov and Grassberger 2009) and as a distance measure (Cover and Thomas 1991). It has been shown that mutual information as a similarity measure is better than the Pearson correlation or Euclidean distance (Priness et al. 2007).

Directional Information Transfer

When comparing objects with different marginal or joint pieces of information, one should preferably use a relative measure rather than an absolute one, so as to minimize the dependence on total information (Kraskov et al. 2005). Hence, mutual information should be standardized to form an index known as DIT. Directional information transfer is the fraction of the information transferred from one site to another. The concept of DIT was introduced by Coombs et al. (1970) in the field of mathematical psychology as

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

a coefficient of constraint (Fass 2006). It is a normalized version of mutual information between two gauges to obtain the fraction of information transferred from one site to another as a value between 0 and 1. Directional information transfer is a much better index than mutual information because the upper bound of mutual information can vary from site to site, depending on the marginal entropy value at the respective station, which makes the mutual information a not so good index of dependence. The DIT can thus be expressed as

$$\mathrm{DIT}_{xy} = \frac{T(X,Y)}{H(X)}; \qquad \mathrm{DIT}_{yx} = \frac{T(X,Y)}{H(Y)} \tag{11}$$

where DIT_{xy} = fractional information inferred by station X about Y; DIT_{yx} = fractional information inferred by station Y about X; $T(X,Y)$ = mutual information between X and Y; and $H(X)$ and $H(Y)$ = marginal entropy values for X and Y, respectively. Because $H(X|Y)$ is equivalent to the loss of information H_{lost}, the formula can be rewritten as

$$\mathrm{DIT} = (H - H_{\mathrm{lost}})/H = 1 - (H_{\mathrm{lost}}|H) \tag{12}$$

It should also be noted that although the mutual information term is symmetric, DIT is no longer symmetric. The concept of using entropy for the purpose of regionalization in hydrology was introduced by Yang and Burn (1994). Alfonso et al. (2010) used DIT as a criterion to determine the independency of water level monitoring stations, which helped in designing an optimum network.

Application of DIT for Regionalization

While using DIT for regionalization, those stations for which both DIT_{xy} and DIT_{yx} are high can be considered to be strongly dependent, because information can be mutually inferred between them. If neither DIT is high, then the two stations should remain in separate groups. If only one DIT is high, say DIT_{xy}, then station Y, whose information can be predicted by X, can join station X if station Y does not belong to any other group; otherwise, it stays in its own group. However, by no means can X enter station Y's group (Yang and Burn 1994). The DIT can be distinguished from traditional similarity measures like correlation coefficient because it is based on the information connection between stations.

The number of groups formed is controlled by the threshold value of DIT. A higher threshold value will lead to a larger number of groups. However, the size of each group will be small. A lower threshold value will result in the formation of a small number of groups, but the size of each group will be larger. There is no rule based on which the threshold of DIT can be fixed, and hence is case specific.

Table 4 shows a sample DIT matrix for eight stations. Say the threshold is considered to be 0.5 in this example. It can be seen that the maximum DIT value corresponds to station pair 2 and 3 (0.52 and 0.50), and the smallest is for station pair 1 and 5 (0.12 and 0.14), respectively. Consider a threshold of 0.35. The groups formed based on the grouping principles explained previously comprise 1, 2, 3, 6, and 8 in groups 1 and 4, and 5 and 7 in group 2. Instead of 0.35, if a lower threshold of for example 0.2 is chosen, then all eight stations will fall under one group. This shows that the lower the threshold, the smaller the group numbers, and the larger the group size. If a higher threshold is chosen, for example 0.45, then it can be seen that initially, stations 1, 2, and 3 fall in one group, and 4 and 7 fall in another group. For stations 5 and 8, there is no combination for which both DIT_{xy} and DIT_{yx} are higher than the threshold. Next, it is checked whether any one value of DIT_{xy} or DIT_{yx} is higher than the threshold. It can be seen that DIT_{18} is 0.47, which is higher than the threshold. Because station 8 does not belong to any group, it can be put into the group of station 1. For station 5, because none of the DIT_{xy} or DIT_{yx} values are above the threshold, it does not fall in either group 1 or 2. Figs. 6(a–c) show the grouping when the threshold is 0.45, 0.35, and 0.2, respectively.

Table 4. Sample DIT Matrix for Eight Stations

Station	1	2	3	4	5	6	7	8
1	1	0.54	0.20	0.13	0.12	0.20	0.21	0.47
2	0.45	1	0.52	0.17	0.15	0.32	0.25	0.32
3	0.25	0.50	1	0.28	0.15	0.49	0.19	0.29
4	0.18	0.15	0.21	1	0.42	0.25	0.48	0.22
5	0.14	0.15	0.17	0.40	1	0.21	0.29	0.15
6	0.22	0.30	0.50	0.22	0.19	1	0.19	0.23
7	0.26	0.27	0.22	0.50	0.31	0.23	1	0.20
8	0.41	0.28	0.26	0.16	0.14	0.20	0.21	1

Once the groups are formed, the criteria of S-DIT may be used for further prioritizing the stations within that group. From the aforementioned examples, consider a threshold of 0.45. As explained previously, three groups will be formed. Consider group 1, which has stations 1, 2, 3, 6, and 8. To prioritize them, calculate S-DIT for each of the stations. The S-DIT for any station i can be calculated as

$$\text{S-DIT}_i = \sum_{j=1, j\neq i}^{N} \mathrm{DIT}_{ij} \tag{13}$$

where N = total number of stations within a region; and DIT_{ij} = information inferred about station j by station i.

Table 5 gives the S-DIT values for all the stations coming under group 1. It can be seen from the table that station 2 should be given highest priority because it has the highest S-DIT and therefore the highest information content among all the stations within that group, and is followed by stations 3, 1, 6, and 8. This criterion will be helpful for station elimination from a group in case a smaller group size is required.

Regional Homogeneity Test

To check the heterogeneity of the regions obtained, the test suggested by Hosking and Wallis (1993, 1997) was performed. This test aims at estimating the degree of heterogeneity among the grouped sites and then assessing whether it is reasonable to treat it as a homogenous region or not. Three heterogeneity measures (HM) were devised, and the values of HM should ideally be less than 1 for the regions to be considered as acceptably homogenous, and between 1 and 2 to be considered as possibly homogenous. If the value of HM is greater than or equal to 2, the region is definitely heterogeneous. The first HM, H_1, is based on the L-coefficient of variation (L-CV), the second HM, H_2, is based on L-CV and L-skewness, and the third measure, H_3, is based on L-skewness and L-kurtosis. They are given as

$$H_1 = \frac{(V - \mu_{v_1})}{\sigma_{v_1}} \qquad H_2 = \frac{(V_1 - \mu_{v_2})}{\sigma_{v_2}} \qquad H_3 = \frac{(V_2 - \mu_{v_3})}{\sigma_{v_3}} \tag{14}$$

$$V = \left[\frac{\sum_{i=1}^{N} n_i (t_i - t_R)^2}{\sum_{i=1}^{N} n_i}\right]^{1/2} \tag{15}$$

J. Hydrol. Eng. 2013.18:870-887.

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

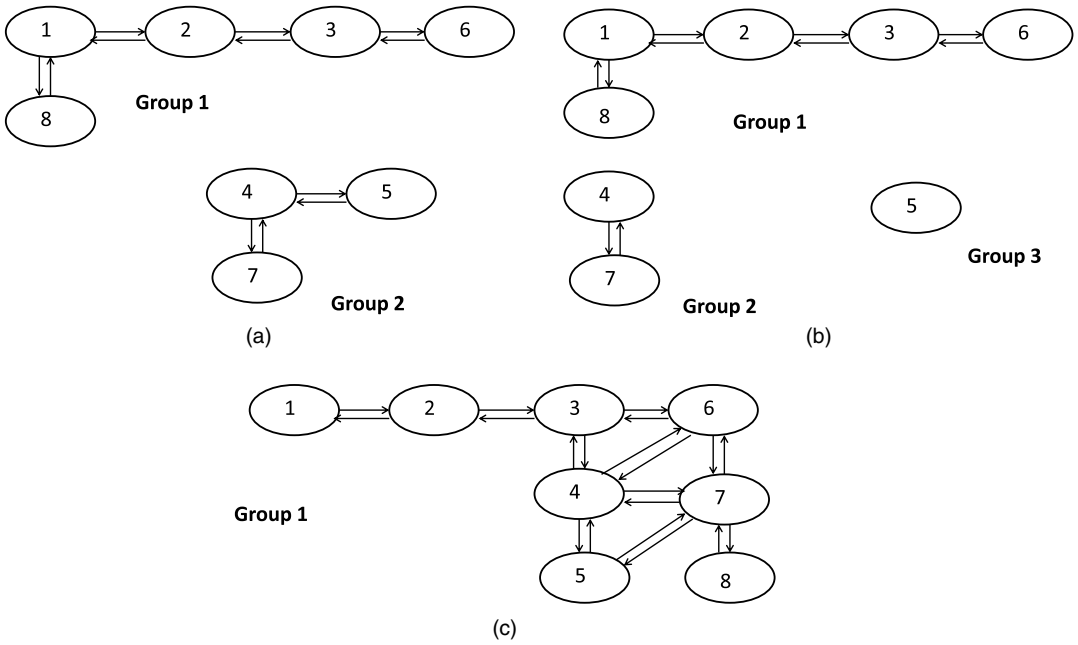


Fig. 6. (a) Grouping pattern for a threshold DIT of 0.35; (b) grouping pattern for a threshold DIT of 0.45; (c) grouping pattern for a threshold DIT of 0.2

Table 5. S-DIT Matrix for Group 1 Stations for Threshold of 0.45

Stations	1	2	3	6	8	S-DIT (summation DIT across the row)
1	1	0.54	0.20	0.20	0.47	1.41
2	0.45	1	0.52	0.32	0.32	1.61
3	0.25	0.50	1	0.49	0.29	1.53
6	0.22	0.30	0.50	1	0.23	1.25
8	0.41	0.28	0.26	0.20	1	1.15

$$V_1 = \left\{ \frac{\sum_{i=1}^{N} n_i[(t_i - t_R)^2 + (t_{3i} - t_{3R})^2]^{1/2}}{\sum_{i=1}^{N} n_i} \right\} \quad (16)$$

$$V_2 = \left\{ \frac{\sum_{i=1}^{N} n_i[(t_{3i} - t_{3R})^2 + (t_{4i} - t_{4R})^2]^{1/2}}{\sum_{i=1}^{N} n_i} \right\} \quad (17)$$

where n_i = record length at the ith grid considered out of a total of N grids; and t_i, t_{3i}, and t_{4i} = L-CV, L-skewness, and L-kurtosis at the respective grid, whereas t_R, t_{3R}, and t_{4R} = weighted average of L-CV, L-skewness, and L-kurtosis, respectively, for the entire region under consideration. Here V, V_1, and V_2 = statistics for the real region; V = weighted standard deviation of L-CVs at the site; V_1 = weighted average distance from the site to the group weighted mean in a two-dimensional space of L-CV and L-skewness; and V_2 = weighted average distance from the site to the group weighted mean in a two-dimensional space of L-skewness and L-kurtosis (Srinivas et al. 2008). The record lengths at the sites were used as the weighting factor. A kappa distribution was then fitted to the regional average of the first four L-moments, and a large number of values were simulated. In this study, the number of simulations (N_{sim}) was chosen as 500, which was considered to be adequate for testing homogeneity (Hosking and Wallis 1997). Each realization constitutes a homogenous region with N sites having the same record length as the real region counterpart. The σ and μ values correspond to the mean and standard deviation of the N_{sim} values of V, V_1, and V_2. Hence, the statistics of the simulated region are compared to the real region. The last two statistics lack power to discriminate homogeneous and heterogeneous regions, and even for grossly heterogeneous regions, they will rarely yield values larger than 2. For the first statistic, the region is considered to be acceptably homogeneous if $H_1 < 1$, possibly homogeneous if H_1 is between 1 and 2, and definitely heterogeneous if H_1 is greater than or equal to 2 (Hosking and Wallis 1997).

The following was suggested by Hosking and Wallis (1997) to improve the homogeneity of the region: (1) elimination or transfer of discordant sites; (2) dividing a region to form two or more new regions; (3) merging two or more regions to form a new region; and (4) obtaining more data and regionalizing again. The discordance measure used for elimination of discordant sites from a region is given by

$$D_i = \frac{1}{3} N(u_i - \bar{u})^T S^{-1}(u_i - \bar{u}) \quad (18)$$

where $u_i = [\, t_i \quad t_{3i} \quad t_{4i} \,]^T$; $\bar{u}$ = average of the L-moment ratios for the region; and S = matrix given by

$$S = \sum_{i=1}^{N} (u_i - \bar{u})(u_i - \bar{u})^T \quad (19)$$

Generally, sites with D-statistics greater than 3 are considered to be discordant from the rest of the region (Hosking and Wallis 1997).

Results and Discussions

Calibration and Validation of VIC Model

Because the VIC model involves a lot of parameters, calibration of the same can become quite tedious. The recommended parameters and the plausible range of values for each of them are given in Table 6. In this study, six soil parameters were considered for calibration purposes. The VIC model calibration was performed using

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

Table 6. Model Parameters for Calibration of VIC Model

Soil parameter	Unit	Range of values
Infiltration shape parameter (b_{inf})	None	0–0.4
Maximum subsurface flow rate (Ds_{max})	mm/day	0–30
Fraction of Ds_{max} when nonlinear flow starts (Ds)	None	0–1
Depth of second soil layer (D_2)	m	0.1–1.5
Depth of third soil layer (D_3)	m	0.1–1.5
Fraction of maximum soil moisture when nonlinear flow starts (W_s)	None	0–1

a random autostart simplex method program. The simplex method was applied using random autostart populations of 75–100 parameter sets. The entire cycle was repeated 5–10 times for each subbasin. Each autostart yielded different R^2 values (usually within ± 0.1) and different parameter sets. As far as the calibration of the routing model was concerned, the suggested parameters for adjustment included velocity and diffusivity. The model developers were less specific about the routing model calibration as compared to the VIC model calibration. Application-based studies focusing on the monthly discharge from large basins have shown that it does not require high accuracy in the routing model parameters. Hence, whereas parameters like flow direction and contributing fraction can be obtained from the DEM, for other parameters like flow velocity and diffusivity, physically reasonable values will be chosen without further calibration (Gao et al. 2010). If only monthly streamflows are required, diffusivity and velocity values of 800 m^2/s and 1.5 m/s are deemed acceptable. In case daily flows are required, the calibration methodology to be followed for routing parameters is outlined in Lohmann et al. (1996, 1998).

The streamflow obtained after calibrating the model parameters was validated using the USGS streamflow data. For this purpose, the routing model was used to route the flow to the selected station locations. Results from the routing model were aggregated to a monthly scale (in cu f/s) and compared with the observed gauge data (in cu f/s). The three performance criteria selected were correlation coefficient, the Nash-Sutcliffe (N-S) efficiency, and mean flow ratio, defined as

Correlation coefficient,

$$r = \frac{M\sum_{i=1}^{M} S_i O_i - \sum_{i=1}^{M} S_i \sum_{i=1}^{M} O_i}{\sqrt{[M\sum_{i=1}^{M} S_i^2 - (\sum_{i=1}^{M} S_i)^2][M\sum_{i=1}^{M} O_i^2 - (\sum_{i=1}^{M} O_i)^2]}} \quad (20)$$

$$\text{Nash Sutcliffe efficiency} = 1.0 - \frac{\sum_{i=1}^{M} (O_i - S_i)^2}{\sum_{i=1}^{M} (O_i - \bar{O})^2} \quad (21)$$

$$\text{Meanflow(MF) ratio} = \frac{\bar{S}}{\bar{O}} \quad (22)$$

where M = number of months; S_i = simulated streamflow for the ith month; O_i = observed streamflow for ith month; and $\bar{S}$ and $\bar{O}$ = mean monthly simulated and observed streamflows, respectively.

A higher value of correlation coefficient and the Nash-Sutcliffe (N-S) efficiency indicate good performance of the model. The closer the value is to 1, the more accurate the model is. Validation of the results obtained from the calibrated model with respect to the observed streamflow values at the respective gauges are shown in Fig. 7. Table 7 gives a summary of performance measures at each of these stations. The validation period was 2 years. The start and end dates of the validation periods for each station are given in Table 1. Because the time period considered in the study was lengthy (1950–2000), different validation periods were considered for the stations such that it covered the time period under consideration. The correlation coefficient values for the 16 stations lie within the range 0.78–0.96, which means the model is capable of explaining 78–96% of variability in the observed data. The N-S efficiency values range from 0.61–0.97. Because an N-S value of 1 corresponds to a perfect match and 0 corresponds to the situation in which simulated values match the mean of observed values, a value of 0.5 may be considered to represent a mediocre model. Hence, from the values obtained for the model at all 16 stations, it can be seen that the model performance is satisfactory. The mean flow ratio values for the model ranges from 0.65–1.81. It can also be seen from Table 7 that the mean flow ratio values are lower than 1 at most of the stations. Thus, the model shows a tendency to underpredict the streamflow values at most of the stations.

Drought Characterization

Table 3 shows the results of the Kolmogorov-Smirnov test conducted to assess the goodness of fit at a 5% significance level for two-parameter lognormal distribution. It can be seen that the test fails to reject the null hypothesis that the values come from two-parameter lognormal distribution. Fig. 5 shows the Q-Q plots for the two-parameter lognormal distribution used to fit streamflow data from the selected stations. From Table 3 and Fig. 5, it can be seen that the fitting of streamflow time series from selected locations within each of the five climatic zones to two-parameter lognormal distribution give a reasonably good fit, and hence this distribution was selected for further calculation of standardized streamflow index. The drought characteristics' severity and duration was obtained using the theory of runs.

Fig. 8 shows the annual average severity, and Fig. 9 shows the maximum drought duration values within each grid for the time period 1950–2000 in the state of Texas. It can be seen that the severity trend matches the precipitation pattern within Texas, indicating an increasing trend in the severity from east to west. The maximum severity was experienced in the western part and went up to 7.64. Similarly, maximum drought durations were also low in the eastern side, whereas the central and western parts showed higher values of duration. The maximum duration had the highest value of 90 months, and the lowest value of 3 months.

Grouping of Grids

There were a total of 4,174 grids of 1/8 degree size that covered the state of Texas. The number of regions formed depended on the threshold value of DIT. Table 8 shows the number of groups formed with the threshold value of DIT varied for drought severity and duration. Because a DIT value higher than 0.5 ensures a good information connection between two grids and higher values yield a large number of groups, eight regions based on drought severity, and nine regions based on drought duration were chosen. The corresponding threshold value of DIT was 0.5 for regions based on severity, and 0.55 for regions based on duration. Once the regions were formed, the next step was to check for their meaningfulness. The L-moments–based heterogeneity test by

J. Hydrol. Eng. 2013.18:870-887.

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

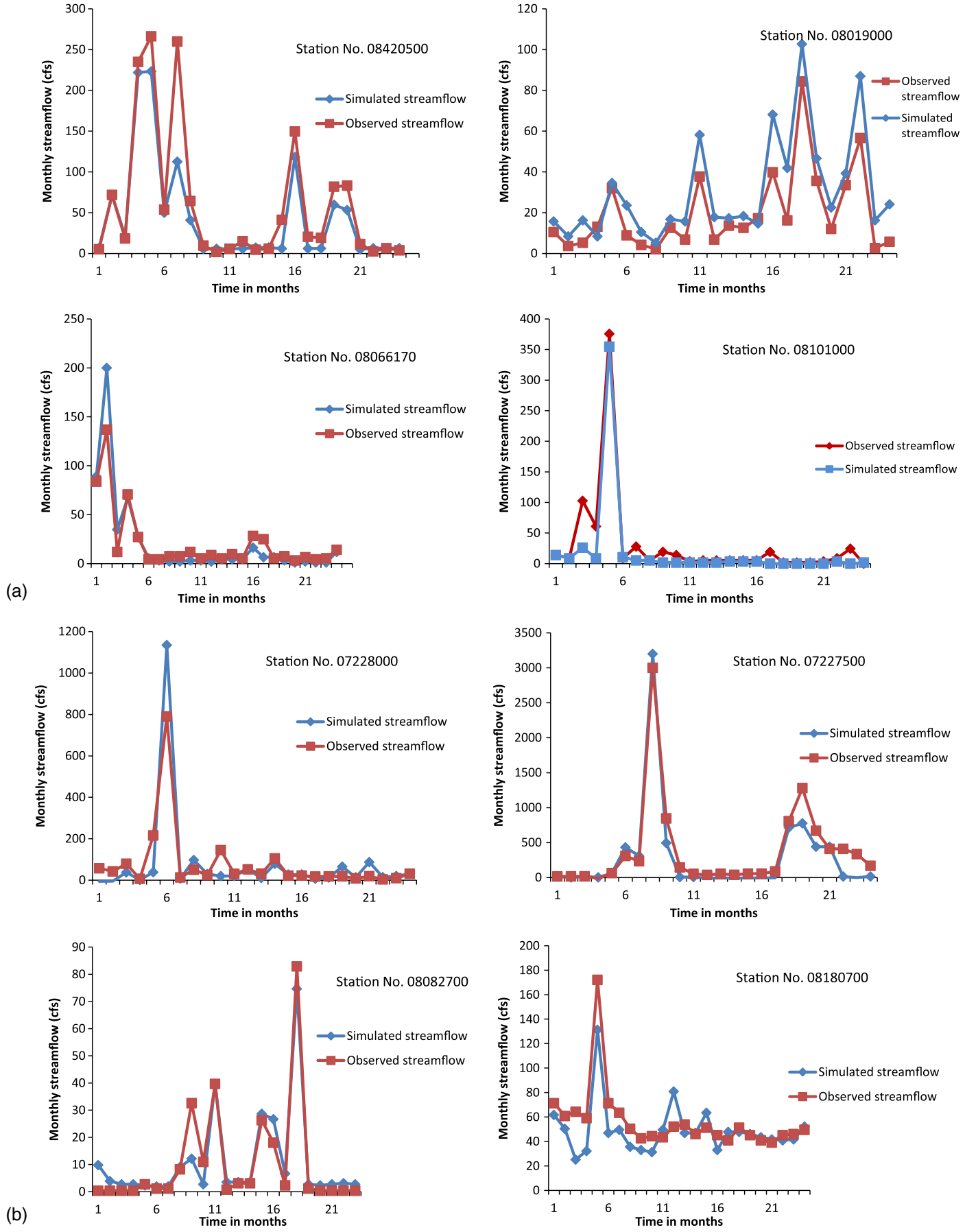


Fig. 7. Comparison of simulated and observed streamflows for selected stations

J. Hydrol. Eng. 2013.18:870-887.

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

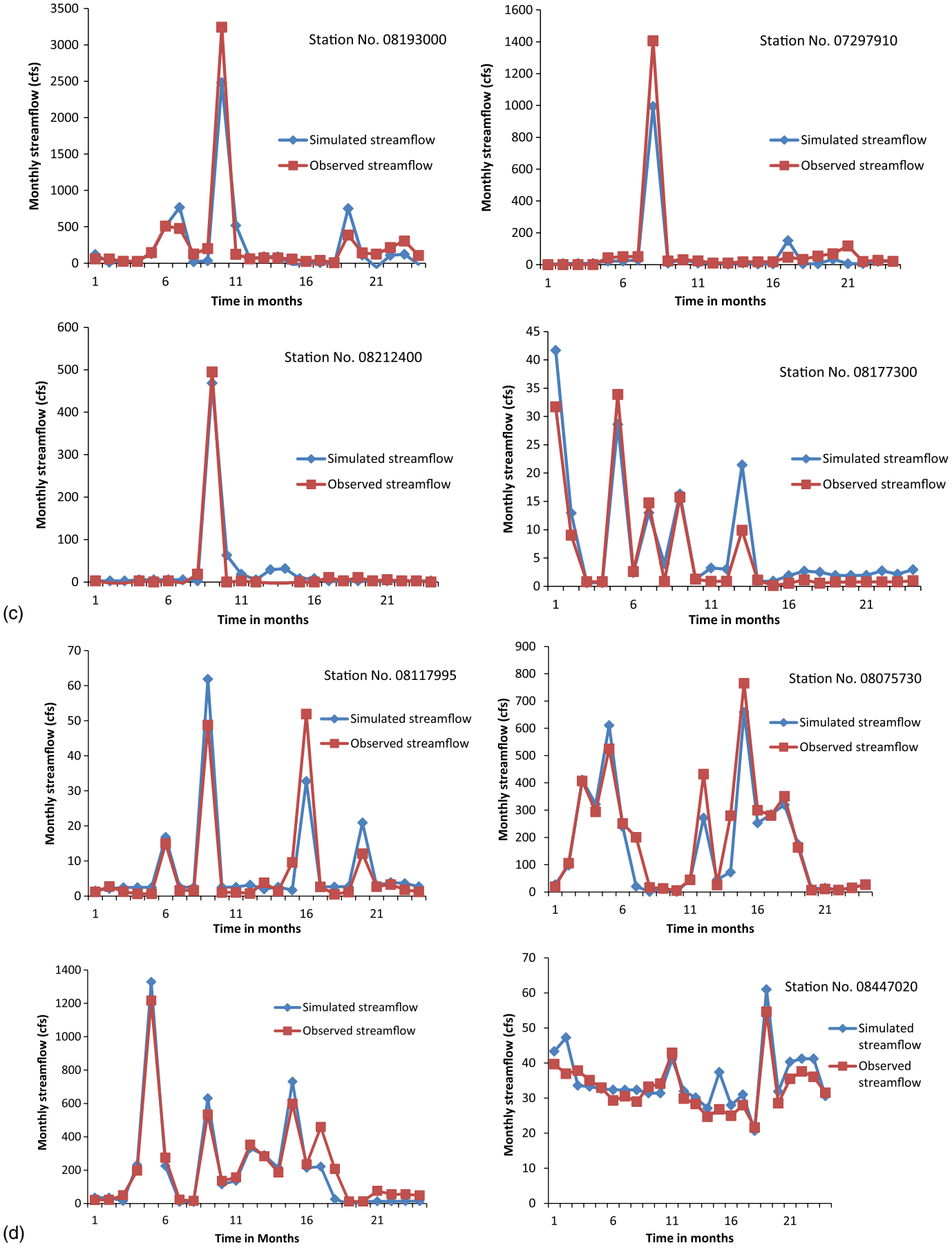


Fig. 7. (*Continued.*)

J. Hydrol. Eng. 2013.18:870-887.

Table 7. Goodness of Fit Test Values of Model Validation at the Selected Stations

Station name	Correlation coefficient	Mean flow ratio	Nash Sutcliffe efficiency
Pecos River at Pecos	0.88	0.71	0.79
Canadian River near Amarillo	0.94	0.77	0.91
Prairie Dog Town fork Red River near Wayside	0.96	0.65	0.89
Canadian River near Canadian	0.89	1.04	0.85
Independence Creek near Sheffield	0.81	1.06	0.62
Nueces River near Asherton	0.89	0.87	0.88
Colorado River near Gail	0.84	0.92	0.64
Colorado River near Stacy	0.94	0.89	0.92
Millers Creek near Munday	0.90	1.09	0.90
Medina River near Macdona	0.81	0.88	0.61
Cowhouse Creek at Pidcoke	0.93	0.68	0.91
Perdido Creek at M 622 near Fannin	0.89	1.35	0.85
Los Olmos Creek near Falfurrias	0.97	1.19	0.97
Lake Fork Creek near Quitman	0.89	0.86	0.87
Kickapoo Creek near Onalaska	0.92	0.82	0.81
Vince Bayou at Pasadena	0.78	1.81	0.82

Hosking and Wallis (1997) was used for this purpose. To improve the homogeneity of a region, the discordant sites within each region were identified by computing a discordance measure. Any station that had a discordant measure value more than 3 was shifted to another region, provided the other region remained homogeneous even after the transfer. If the aforementioned condition was not satisfied, a site could not be allocated to any other region, and hence it would be eliminated. Tables 9 and 10 give details of the discordant sites within the regions formed based on DIT for drought severity and duration, respectively. Tables 11 and 12 show the heterogeneity measures for the regions after elimination or shifting of discordant sites. A total of eight regions were formed based on drought severity, and nine regions were formed based on drought duration. From Table 11, which shows the measures for regions based on drought severity, it can be seen that all three heterogeneity measures given by Hosking and Wallis (1997) are less than 1 in all cases except region 2, and hence they can be considered to be acceptably homogeneous. In the case of region 2, because measure H_3 is more than 1, the region can be considered possibly homogeneous. Table 12 shows the heterogeneity measures for the regions based on drought duration. All nine regions had heterogeneity measures less than 1, and hence the regions can be considered to be acceptably homogeneous. Figs. 10 and 11 show the homogenous regions formed based on the drought severity and drought duration, respectively. Tables 13 and 14 give details of the regions based on severity and duration, respectively. Fig. 12 serves as a reference for the zones in Texas based on the differences in topography, climate, and environment, which will be used in further discussions. It can be seen from Table 13 that region 1 is the most critical zone in terms of drought severity. The average drought severity over this region comes out to be 7.65. Region 1 covers approximately 11.5% of the area of Texas and lies in the Trans Pecos zone of Texas. Region 8, which lies within parts of lower valley and upper coast, is the region that is least severe. The average severity of this region comes out to be 4.898. Region 3, which lies within the Edwards Plateau and Low Rolling Plains zones, is the largest in area, and covers approximately 15.8% of the state of Texas. The average severity for this region comes to 6.294. Region 7, which lies within

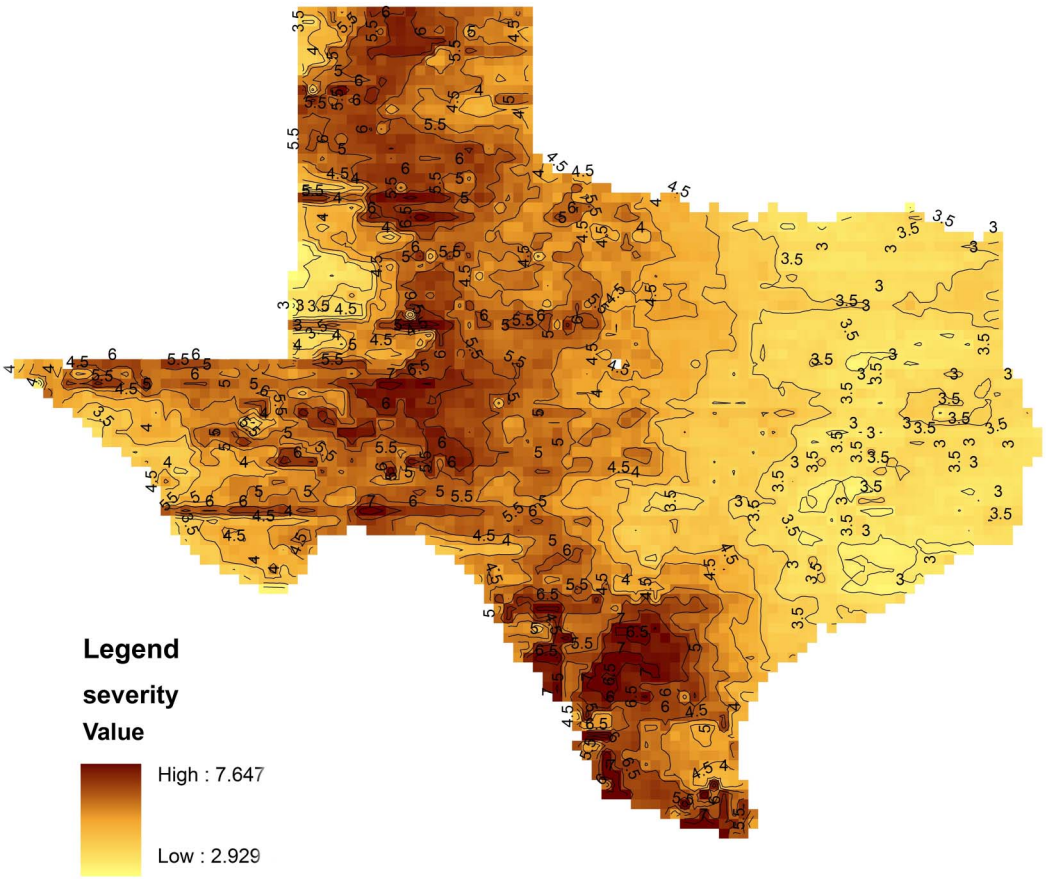


Fig. 8. Annual average drought severity pattern for Texas during 1950–2000

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

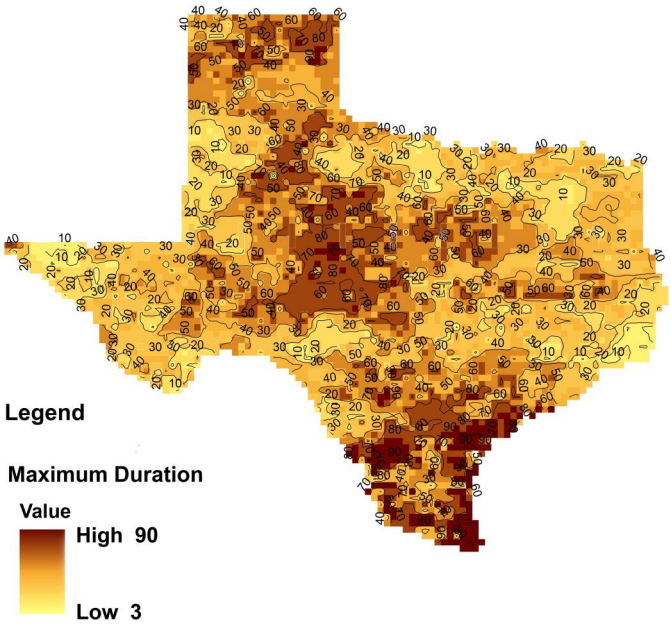


Fig. 9. Maximum drought duration pattern for Texas during 1950–2000

Table 8. Number of Regions Formed by Varying Thresholds

Drought severity		Drought duration	
Threshold DIT	Number of regions	Threshold DIT	Number of regions
0.2	4	0.15	3
0.25	5	0.3	5
0.35	7	0.45	6
0.5	8	0.55	9

Table 9. Discordant Sites in the Regions Formed based on Drought Severity

Region	Number of discordant sites	Adjustments
Region 1	8	4 deleted 4 moved to region 4
Region 2	7	3 deleted 4 moved to region 3
Region 4	0	—
Region 5	0	—
Region 6	9	2 moved to region 4 1 moved to region 5 6 moved to region 7
Region 7	12	4 deleted 4 moved to region 4 1 moved to region 6 3 moved to region 8
Region 8	8	4 deleted 2 moved to region 5 2 moved to region 6

Table 10. Discordant Sites in the Regions Formed based on Drought Duration

Region	Number of discordant sites	Adjustments
Region 1	9	5 deleted 1 moved to region 2 1 moved to region 3 2 moved to region 5
Region 2	4	4 deleted
Region 3	0	—
Region 4	13	7 deleted 1 moved to region 3 3 moved to region 5 2 moved to region 8
Region 5	5	3 deleted 2 moved to region 1
Region 6	5	2 deleted 1 moved to region 5 2 moved to region 8
Region 7	6	2 deleted 2 moved to region 8 2 moved to region 9
Region 8	4	2 deleted 2 moved to region 6
Region 9	0	—

Table 11. Heterogeneity Measures for the Regions based on Drought Severity

Region	H_1	H_2	H_3	Conclusion
Region 1	−1.03	−1.68	0.352	Acceptably homogeneous
Region 2	−1.299	−3.09	1.14	Possibly homogeneous
Region 3	−1.332	0.376	0.241	Acceptably homogeneous
Region 4	−1.325	−1.698	0.189	Acceptably homogeneous
Region 5	−1.670	−8.703	−1.658	Acceptably homogeneous
Region 6	−2.176	−7.469	0.924	Acceptably homogeneous
Region 7	−1.481	−1.125	−1.636	Acceptably homogeneous
Region 8	−1.346	−1.008	−1.475	Acceptably homogeneous

Table 12. Heterogeneity Measures for the Regions based on Drought Duration

Region	H_1	H_2	H_3	Conclusion
Region 1	−2.514	0.894	0.722	Acceptably homogeneous
Region 2	−2.159	0.935	0.639	Acceptably homogeneous
Region 3	−2.682	0.946	0.644	Acceptably homogeneous
Region 4	−3.034	0.575	0.477	Acceptably homogeneous
Region 5	−2.477	−3.520	−2.205	Acceptably homogeneous
Region 6	−2.176	−7.469	0.924	Acceptably homogeneous
Region 7	−1.481	−1.125	−1.636	Acceptably homogeneous
Region 8	−1.162	−5.728	−4.716	Acceptably homogeneous
Region 9	−2.265	0.355	0.983	Acceptably homogeneous

south Texas, south central Texas, and the lower valley, is the smallest in area, and covers approximately 10.9% of the state of Texas. The average severity of region 7 comes to be 5.346. From Fig. 10 and Table 14, it can be seen that the region with the longest drought duration is region 6, which lies in south Texas, south central Texas, and the lower valley, and covers 11.69% of Texas. The average longest duration within region 6 is 91 months. Region 9, which lies within the upper coast and east Texas zones and covers approximately 11.7% of Texas, has an average longest duration of 27 months, which is the least among all the regions. Fig. 13 shows the patterns of various drought categories in Texas based on the classification given in Table 2. It can be seen that only a few regions—regions 1, 2, 6, 7, and 8 (refer to Fig. 10 for identification of regions) lying in the Trans Pecos, High Plains, Upper Coast,

J. Hydrol. Eng. 2013.18:870-887.

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

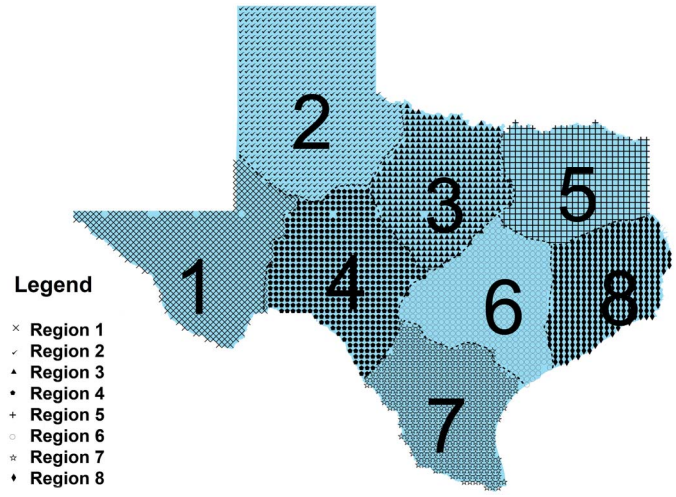


Fig. 10. Homogenous regions formed using DIT based on drought severity

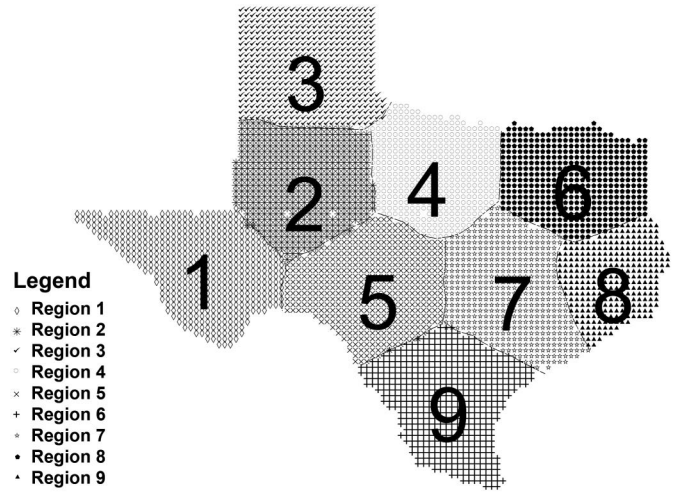


Fig. 11. Homogeneous regions formed based on drought duration

Table 13. Details of the Regions Formed based on Drought Severity

Region	Number of grids	Percentage area covered	Annual average severity
1	478	11.495	7.65
2	489	11.761	7.219
3	658	15.824	6.294
4	574	13.804	6.632
5	550	13.227	7.074
6	483	11.616	5.435
7	453	10.895	5.346
8	473	11.375	4.898

north, and south central Texas—suffer from moderately dry droughts. Most parts of the Trans Pecos, High Plains, Edwards Plateau, Low Rolling Plains, south Texas, and lower valley regions are affected by severely dry to extremely dry droughts, whereas the eastern part of Texas is not affected by severe or extremely dry droughts.

Table 14. Details of Homogenous Regions Formed based on Drought Duration

Region	Number of grids	Percentage area covered	Average drought duration in months
1	499	11.11	73
2	462	10.52	64
3	498	12.02	58
4	485	9.13	47
5	484	10.51	77
6	436	11.69	91
7	437	11.37	33
8	473	12.01	42
9	379	11.66	27

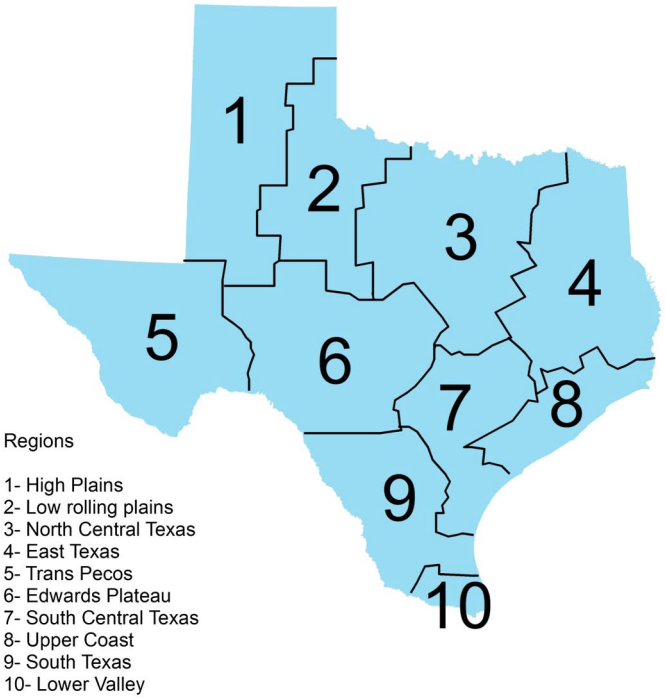


Fig. 12. Zones in Texas based on topography and climate

The study that has been carried out for the univariate case can be extended to the bivariate case by considering the drought severity and duration together while the homogeneous regions are formed. The procedure followed for the bivariate case will be similar to the univariate case. However, the calculation of joint probabilities will be more complicated because a four-dimensional contingency table would be required for the same. The threshold value considered was 0.4, corresponding to a total of five homogeneous regions that were formed for the bivariate case. Table 15 gives the details of the heterogeneity measures for the five regions formed for the bivariate case. Fig. 14 shows the homogeneous regions formed for the bivariate case. The details of each of the regions are given in Table 16. It can be seen from Table 16 that region 5, which covers 24.23% of the area of Texas, has the minimum severity and duration. This region falls under the zone of east Texas and parts of the upper coast and north central Texas. Region 2, which covers 19.2% of the area of Texas, has the maximum severity, and comes under the Trans Pecos zone. Region 1, which covers 18.45% of the area of Texas, has the maximum duration, and comes under the lower valley, south Texas, and parts of the south central Texas zones.

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

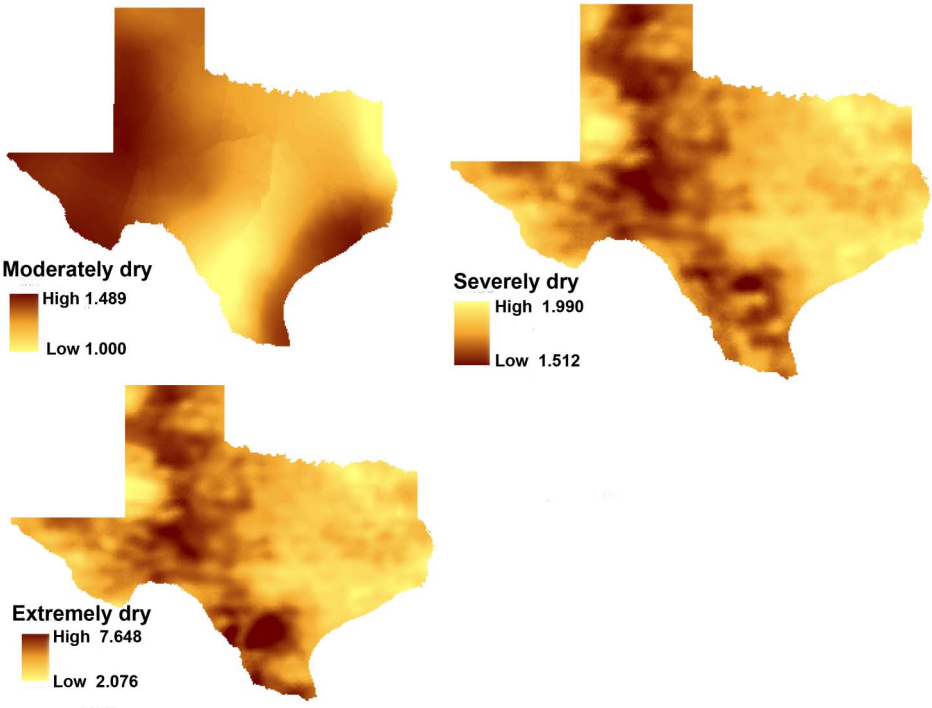


Fig. 13. Spatial pattern of drought categories within Texas

Table 15. Heterogeneity Measures for the Regions Formed in the Bivariate Case

Region	H_1	H_2	H_3	Conclusion
Region 1	0.839	1.023	0.764	Possibly homogeneous
Region 2	−1.173	0.927	0.873	Acceptably homogeneous
Region 3	0.026	0.583	0.698	Acceptably homogeneous
Region 4	0.905	0.884	0.329	Acceptably homogeneous
Region 5	0.926	1.183	0.547	Possibly homogeneous

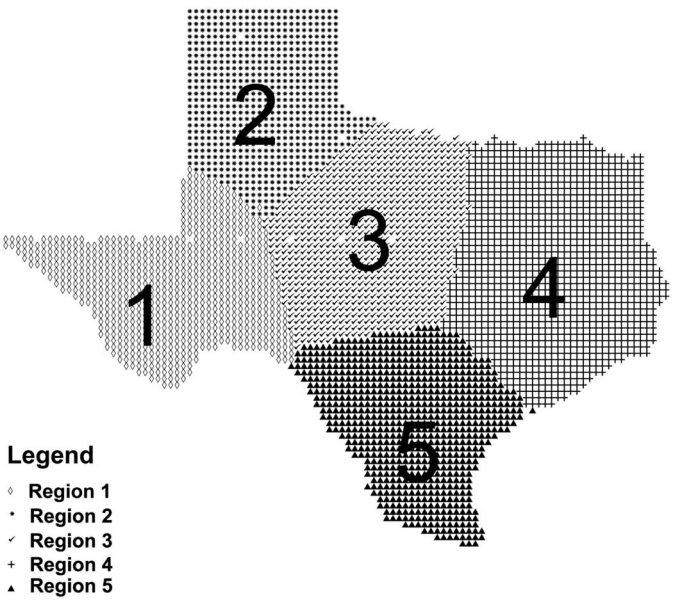


Fig. 14. Homogeneous regions formed in the bivariate case

Table 16. Details of Homogenous Regions Formed based on Drought Severity and Duration

Region	Number of grids	Percentage area covered	Annual average severity	Average drought duration in months
1	759	18.45	6.714	87
2	790	19.20	7.937	75
3	694	16.87	6.717	51
4	874	21.24	7.169	54
5	997	24.23	4.814	30

Conclusions

An entropy-based similarity measure known as DIT was used to regionalize the state of Texas based on drought severity and duration. This measure, being more sensitive to nonlinear dependencies, is a better similarity measure than the commonly used linear dependence measures. By making use of the nonsymmetric property of the index, if there is a great difference between the DIT_{xy} and DIT_{yx} values of a station pair, it can imply that given the observations at one station, the response at the other station is ambiguous. This can be because of greater loss during information transfer. It should, however, be noted that no strict guidelines are available for fixing a threshold value for the DIT. This is expected, because regionalization is essentially a subjective process, and hence in any case, the threshold will be user defined provided that the value is not too high (which may lead to strong dependence between stations belonging to different regions) or too low (which may lead to low dependence between stations within the same region). Finer adjustments can be made to the threshold value by observing how a change in its value affects the number and size of the regions formed. The following conclusions were drawn from the study:

1. Directional information transfer can satisfactorily identify homogeneous regions based on drought severity and duration, thus leading to classification of Texas into zones based on streamflow drought properties.

J. Hydrol. Eng. 2013.18:870-887.

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

2. Identification of critical regions in a drought prone state like Texas is done by assessing drought properties within each region formed. Region 1, lying within the Trans Pecos zone in west Texas, is the most critical region in terms of severity. Region 8, which lies in the eastern part of Texas, has the lowest severity. The pattern is consistent with the precipitation pattern in Texas. As far as drought duration is concerned, region 6, which lies in south Texas, south central Texas, and the lower valley, has the longest drought duration. Region 9, which lies in eastern Texas, has the lowest drought duration.
3. Parts of the high plains, upper coast, central, and western Texas are affected by moderately dry droughts. However, severely dry and extremely dry droughts are mainly restricted to western, central, and south Texas.

The study can be extended to the bivariate case too. The streamflows at the USGS gauges are controlled/observed flow. The model has been calibrated and validated on the basis of the original controlled/observed flow instead of naturalized flow. This might have an impact on the runoff simulations obtained from the model. It might not be possible for any model to accurately reproduce the real-world scenario for the runoff production process. However, it should be noted that the model simulations do show a satisfactory correlation with the original streamflow, and that the model in general underpredicts streamflow values in comparison to the original values. As an extension to the present work, it would be interesting to analyze how well the model simulations match if naturalized streamflow values were used instead of controlled flows.

Having obtained the homogeneous zones for streamflow drought with the knowledge of variation of drought properties within each of these regions, a mitigation plan specific to that region can be developed. This will help water resources planners to overcome the gravity of water crisis in coming years.

Acknowledgments

This work has been supported by the USGS project grant 2009TX334G.

References

Abdulla, F. A., Lettenmaier, D. P., Wood, E. F., and Smith, J. A. (1996). "Application of a macroscale hydrologic model to estimate the water balance of the Arkansas-Red river basin." *J. Geophys. Res.*, 101(D3), 7449–7459.

Abramowitz, M., and Stegun, I. A. (1964). "Handbook of mathematical functions." Chapter 26, *Applied mathematics series*, National Bureau of Standards, New York, NY, 931–933.

Acreman, M. C., and Wiltshire, S. E. (1989). "The regions are dead: Long live the regions." Chapter 2, *FRIENDS in hydrology*, No. 187, L. Roald, K. Nordseth, and S. K. A. Hassel, eds., IAHS, Bolkesjo, Norway, 175–188.

Alfonso, L., Lobbrecht, A., and Price, R. (2010). "Optimization of water level monitoring network in polder systems using information theory." *Water Resour. Res.*, 46(12), W12553.

Andreadis, K., and Lettenmaier, D. P. (2006). "Trends in 20th century drought over the continental United States." *Geophys. Res. Lett.*, 33(10), L10403.

Andreadis, K., Wood, C. E., Wood, A., Hamlet, A., and Lettenmaier, D. (2005). "Twentieth century drought in conterminous United States." *J. Hydro. Meteor.*, 6(6), 985–1001.

Benke, A. C., and Cushing, C. E., eds. (2005). *Rivers of North America*, Elsevier Academic, Burlington, MA.

Bhaskar, N. R., and O'Connor, C. A. (1989). "Comparison of method of residuals and cluster analysis for flood regionalization." *J. Water Resour. Plann. Manage.*, 115(5), 567–582.

Bobee, B., and Rasmussen, P. (1995). "Recent advances in flood frequency analysis." *Rev. Geophys.*, 33(S2), 1111–1116.

Bowling, L. C., Lettenmaier, D. P., Njissen, B., Polcher, J., Koster, R. D., and Lohmann, D. (2003a). "Simulation of high latitude hydrological processes in the Torne-Kalix basin: PILPS Phase 2(e) 3: Equivalent model representation and sensitivity experiments." *Global Planet. Change*, 38(1–2), 55–71.

Bowling, L. C., et al. (2003b). "Simulation of high-latitude hydrological processes in the Torne-Kalix basin: PILPS Phase 2(e) 1: Experiment description and summary intercomparisons." *Global Planet. Change*, 38(1–2), 1–30.

Bureau of Economic Geology. (1996). "River basin map of Texas." Univ. of Texas, Austin, TX.

Burn, D. H., and Goel, N. K. (2000). "The formation of groups for regional flood frequency analysis." *Hydrol. Sci. J.*, 45(1), 97–112.

Burn, D. H., Zrinji, Z., and Kowalchuk, M. (1997). "Regionalization of catchments for regional flood frequency analysis." *J. Hydrol. Eng.*, 2(2), 76–82.

Byzedi, M., and Saghafian, B. (2009). "Regional analysis of streamflow drought: A case study for Southwestern Iran." *W. Acad. Sci.*, 57(33), 447–451.

Chokmani, K., and Ouarda, T. B. M. J. (2004). "Physiographical space based kriging for regional flood frequency estimation at ungauged sites." *Water Resour. Res.*, 40(12), W12514.

Choquette, A. F. (1988). "Regionalization of peak discharges for streams in Kentucky." *Water Resources Investigation Rep. 87-4209*, USGS, Louisville District, Louisville, KY.

Clausen, B., and Pearson, C. P. (1995). "Regional frequency analysis of annual maximum streamflow drought." *J. Hydrol.*, 173(1–4), 111–130.

Coombs, C. H., Dawes, R. M., and Tversky, A. (1970). *Mathematical psychology: An elementary introduction*, Prentice-Hall, Oxford, England.

Cosby, B. J., Hornberger, G. M., Clapp, R. B., and Ginn, T. R. (1984). "A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils." *Water Resour. Res.*, 20(6), 682–690.

Cover, T. M., and Thomas, J. A. (1991). *Elements of information theory*, Wiley, New York.

Dickey, D. A., and Fuller, W. A. (1979). "Distribution of the estimators for autoregressive time series with a unit root." *J. Am. Stat. Assoc.*, 74(366), 423–431.

Dunn, R. S. (2011). "Droughts." Handbook of Texas ⟨http://www.tshaonline.org/handbook/online/articles/ybd01⟩.

Fass, D. M. (2006). "Human sensitivity to mutual information." Ph.D. dissertation, Rutgers State Univ., New Brunswick, NJ.

Gao, H., et al. (2010). ⟨http://www.hydro.washington.edu/Lettenmaier/Models/VIC⟩.

Guerrero-Salazar, P., and Yevjevich, V. (1975) "Analysis of drought characteristics by the theory of runs." *Hydrology Paper No. 80*, Colorado State Univ., Fort Collins, CO.

Hansen, M. C., DeFries, R. S., and Townshend, J. R. G., and Sohlberg, R. (2000). "Global land cover classification at 1 km spatial resolution using a classification tree approach." *Int. J. Remote Sens.*, 21(6–7), 1331–1364.

Hisdal, H., and Tallaksen, L. M. (2003). "Estimation of regional meteorological and hydrological drought characteristics: A case study for Denmark." *J. Hydrol.*, 281(3), 230–247.

Hosking, J. R. M., and Wallis, J. R. (1993). "Some statistics useful in regional frequency analysis." *Water Resour. Res.*, 29(2), 271–281.

Hosking, J. R. M., and Wallis, J. R. (1997). "Regional frequency analysis: An approach based on L-moments." Cambridge University Press, New York.

Isik, S., and Singh, V. P. (2008). "Hydrologic regionalisation of watersheds in Turkey." *J. Hydrol. Eng.*, 13(9), 824–834.

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

Kraskov, A., and Grassberger, P. (2009). "MIC: Mutual information based hierarchical clustering." Chapter 5, *Information theory and statistical learning*, F. E. Streib, and M. Dehmer, eds., Springer, New York, 101–123.

Kraskov, A., Stogbauer, H., Andrzejak, R. G., and Grassberger, P. (2005). "Hierarchical clustering using mutual information." *Europhys. Lett.*, 70(2), 278–284.

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

Downloaded from ascelibrary.org by University Of Central Florida Libraries on 10/03/13. Copyright ASCE. For personal use only; all rights reserved.

Kroll, C. N., and Vogel, R. M. (2002). "Probability distribution of low streamflow series in the United States." *J. Hydrol.*, 7(2), 137–146.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., and Shin, Y. (1992). "Testing the null of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" *J. Econometrics*, 54(1–3), 159–178.

Lathi, B. P. (1968). *An introduction to random signals and information theory*, International Textbook Company, Scranton, PA.

Liang, X., Lettenmaier, D. P., and Wood, E. F. (1996). "One-dimensional statistical dynamic representation of subgrid spatial variability of precipitation in the two-layer variable infiltration capacity model." *J. Geophys. Res.*, 101(D16), 403–422.

Liang, X., Lettenmaier, D. P., Wood, E. F., and Burges, S. J. (1994). "A simple hydrologically based model of land surface water and energy fluxes for GSMs." *J. Geophys. Res.*, 99(D7), 14415–14428.

Lohmann, D., Nolte-Holube, R., and Raschke, E. (1996). "A large-scale horizontal routing model to be coupled to land surface parametrization schemes." *Tellus Ser. A: Dyn. Met. Oceanogr.*, 48(5), 708–721.

Lohmann, D., Raschke, E., Nijssen, B., and Lettenmaier, D. P. (1998). "Regional scale hydrology: I. Formulation of the VIC-2L model coupled to a routing model." *Hydrol. Sci. J.*, 43(1), 131–141.

Maurer, E. P., Nijssen, B., and Lettenmaier, D. P. (2000). "Use of reanalysis land surface water budget variables in hydrologic studies." *GEWEX News*, 10(4), 6–8.

Maurer, E. P., Wood, A. W., Adam, J. C., Lettenmaier, D. P., and Nijssen, B. (2002). "A long-term hydrologically-based data set of land surface fluxes and states for the conterminous United States." *J. Clim.*, 15(22), 3237–3251.

McKee, T. B., Doesken, N. J., and Kleist, J. (1993). "The relationship of drought frequency and duration to time scales." *Proc., 8th Conf. App. Clim.*, American Meterological Society, Boston, MA, 179–184.

McMahon, T. A., Pegram, G. G. S., and Vogel, R. M. (2007). "Revisiting reservoir storage yield relationships using a global streamflow database." *Adv. Water Resour.*, 30(8), 1858–1872.

Millan, J., and Yevjevich, V. (1971). "Probabilities of observed droughts." *Hydrology Paper No. 50*, Colorado State Univ., Fort Collins, CO.

Miller, D. A., and White, R. A. (1998). "A conterminous United States multilayer soil characteristics dataset for regional climate and hydrology modeling." *Earth Interact.*, 2(2), 1–26.

Milly, P. C. D., et al. (2008). "Climate change: Stationarity is dead: Whither water management?" *Science*, 319(5863), 573–574.

Mirakbari, M., Ganji, A., and Fallah, S. R. (2010). "Regional bivariate frequency analysis of meteorological drought." *J. Hydrol.*, 15(12), 985–1000.

Mishra, A. K., and Singh, V. P. (2009). "Analysis of drought severity-area-frequency curves using a general circulation model and scenario uncertainty." *J. Geophys. Res.*, 114(D6), D06120.

Mishra, A. K., and Singh, V. P. (2010). "A review of drought concepts." *J. Hydrol.*, 391(1–2), 202–216.

Mitchell, K., et al. (1999). "GCIP land data assimilation system (LDAS) project now underway." *GEWEX News*, 9(4), 3–6.

Modarres, R. (2007). "Streamflow drought time series forecasting." *Stoch. Environ. Res. Risk Assess.*, 21(3), 223–233.

Myneni, R. B., Nemani, R. R., and Running, S. W. (1997). "Estimation of global leaf area index and absorbed PAR using radiative transfer models." *IEEE Trans. Geosci. Remote Sens.*, 35(6), 1380–1393.

Nalbantis, L., and Tsakiris, G. (2009). "Assessment of hydrologic drought revisited." *Water Res. Manage.*, 23(5), 881–897.

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

Nijssen, B. N., Lettenmaier, D. P., Liang, X., Wetzel, S. W., and Wood, E. F. (1997). "Streamflow simulation for continental-scale river basins." *Water Resour. Res.*, 33(4), 711–724.

Nijssen, B. N., O'Donnell, G. M., Lettenmaier, D. P., and Wood, E. F. (2001). "Predicting the discharge of global rivers." *J. Clim.*, 14(15), 3307–3323.

Priness, I., Maimon, O., and Ben-Gal, I. (2007). "Evaluation of gene-expression clustering via mutual information distance measure." *BioInformatics*, 8(111), 1–12.

Rao, A. R., and Srinivas, V. V. (2006a). "Regionalization of watersheds by fuzzy cluster analysis." *J. Hydrol.*, 318(1–4), 57–59.

Rao, A. R., and Srinivas, V. V. (2006b). "Regionalization of watersheds by hybrid cluster analysis." *J. Hydrol.*, 318(1–4), 37–56.

Rawls, W. J., Ahuja, L. R., Brakensiek, D. L., and Shirmohammadi, A. (1993). "Infiltration and soil water movement." *Handbook of hydrology*, D. Maidment, ed., McGraw-Hill, New York, 5.1–5.51.

Reynolds, C. A., Jackson, T. J., and Rawls, W. J. (2000). "Estimating soil water-holding capacities by linking the Food and Agriculture Organization soil map of the world with global pedon databases and continuous pedotransfer functions." *Water Resour. Res.*, 36(12), 3653–3662.

Rodriguez-Iturbe, I. (1969). "Applications of theory of runs to hydrology." *Water Resour. Res. Lett.*, 5(6), 1422–1426.

Salathe, E. P. (2003). "Comparison of various precipitation downscaling methods for the simulation of stream flow in a rain shadow river basin." *Int. J. Clim.*, 23(8), 887–901.

Saldarriaga, J., and Yevjevich, V. (1970). "Application of run-lengths to hydrologic series." *Hydrology Paper No. 40*, Colorado State Univ., Fort Collins, CO.

Satyanarayana, P., and Srinivas, V. V. (2011). "Regionalization of precipitation in data sparse areas using large scale atmospheric variables—A fuzzy clustering approach." *J. Hydrol.*, 405(3–4), 462–473.

Sen, Z. (1976). "Wet and dry periods of annual flow series." *J. Hydraul. Div.*, 102(HY10), 1503–1514.

Sen, Z. (1977). "Run-sums of annual flow series." *J. Hydrol.*, 35(3–4), 311–324.

Sen, Z. (1980). "Regional drought and flood frequency analysis, theoretical consideration." *J. Hydrol.*, 46(3), 251–263.

Shannon, C. E. (1948). "A mathematical theory of communication." *Bell Syst. Tech. J.*, 27(3), 379–423.

Sheffield, J., Goteti, G., Wen, F., and Wood, E. F. (2004). "A simulated soil moisture based drought analysis for the United States." *J. Geophys. Res.*, 109(D24), D24108.

Sheffield, J., and Wood, E. F. (2008). "Global trends and variability in soil moisture and drought characteristics, 1950–2000, from observation-driven simulations of the terrestrial hydrologic cycle." *J. Clim.*, 21(3), 432–458.

Shukla, S., and Wood, A. W. (2008). "Use of a standardized runoff index for characterizing hydrologic drought." *Geophys. Res. Lett.*, 35(2), L02405.

Singh, K. K., and Singh, S. V. (1996). "Space time variation of regionalization or seasonal and monthly summer monsoon rainfall of the sub-Himalayan region and Gangetic plains of India." *Clim. Res.*, 6(3), 251–262.

Srinivas, V. V., Tripathi, S., Rao, A. R., and Govindaraju, R. S. (2008). "Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering." *J. Hydrol. (Amsterdam)*, 348, 148–166.

Sturges, H. (1926). "The choice of a class-interval." *J. Am. Stat. Assoc.*, 21(153), 65–66.

Wilhite, D. A., ed. (2000). "Drought as a natural hazard: Concepts and definitions." Chapter 1, *Drought: A global assessment, hazards disasters series*, Vol. 1, Routledge, New York, 3–18.

Wilhite, D. A., and Glantz, M. H. (1985). "Understanding the drought phenomenon: The role of definitions." *Water Int.*, 10(3), 111–120.

Yang, Y., and Burn, D. H. (1994). "An entropy approach to data collection network design." *J. Hydrol.*, 157(1), 307–324.

Yevjevich, V., Siddiqui, M. M., and Downer, R. N. (1967). "Application of runs to hydrologic droughts." *Proc., Int. Hydrology Symp.*, Vol. 1, Paper 63, Fort Collins, CO, 496–505.

Zaidman, M. D., Ress, H. G., and Young, A. R. (2001). "Spatio-temporal development of streamflow droughts in north-west Europe." *Hydrol. Earth Syst. Sci.*, 6(4), 733–751.

Zrinji, Z., and Burn, D. H. (1994). "Flood frequency analysis or ungauged sites using a region of influence approach." *J. Hydrol.*, 153(1–4), 1–21.

Zrinji, Z., and Burn, D. H. (1996). "Regional flood frequency with hierarchical region of influence." *J. Water Resour. Plann. Manage.*, 122(4), 245–252.