Journal of Hydrology 526 (2015) 164–182

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Multivariate drought index: An information theory based approach for integrated drought assessment

Deepthi Rajsekhar [a,*], Vijay. P. Singh [b,a], Ashok. K. Mishra [c]

[a] Biological and Agricultural Engineering (BAEN) Department, Texas A & M University, College Station, TX 77840, United States
[b] Civil Engineering Department, Texas A & M University, College Station, TX 77840, United States
[c] Glenn Department of Civil Engineering, Clemson University, Clemson, SC 29634, United States

ARTICLE INFO

Article history:
Available online 26 November 2014

Keywords:
Entropy
Kernel ECA
Kernel PCA
Multivariate drought index

SUMMARY

Most of the existing drought indices are based on a single variable (e.g. precipitation) or a combination of two variables (e.g., precipitation and streamflow). This may not be sufficient for reliable quantification of the existing drought condition. It is possible that a region might be experiencing only a single type of drought at times, but multiple drought types affecting a region is quite common too. To have a comprehensive representation, it is better to consider all the variables that lead to different physical forms of drought, such as meteorological, hydrological, and agricultural droughts. Therefore, we propose to develop a multivariate drought index (MDI) that will utilize information from hydroclimatic variables, including precipitation, runoff, evapotranspiration and soil moisture as indicator variables, thus accounting for all the physical forms of drought. The entropy theory was utilized to develop this proposed index, that led to the smallest set of features maximally preserving the information of the input data set. MDI was then compared with the Palmer drought severity index (PDSI) for all climate regions within Texas for the time period 1950–2012, with particular attention to the two major drought occurrences in Texas, viz. the droughts which occurred in 1950–1957, and 2010–2011. The proposed MDI was found to represent drought conditions well, due to its multivariate, multi scalar, and nonlinear properties. To help the user choose the right time scale for further analysis, entropy maps of MDI at different time scales were used as a guideline. The MDI time scale that has the highest entropy value may be chosen, since a higher entropy indicates a higher information content.

1. Introduction

Droughts are the costliest of all natural disasters with an estimated annual loss of $6–8 billion in the United States (Wilhite, 2000) and collectively affects more people than any other natural disaster. Thus, there is a need for developing a system to quantify, monitor and predict droughts (Mishra and Singh, 2011). However, given the wide variety of sectors affected by drought and its diverse geographical and temporal distribution, it is difficult to develop a single, precise definition for drought.

Droughts are classified into four categories: meteorological or climatological, agricultural, hydrological, and socioeconomic (The American Meteorological Society, 2004; Mishra and Singh, 2010). A prolonged deficit in precipitation leads to meteorological drought. A dryness in the surface layers (root zone), which occurs at a critical time during the growing season, can result in an agricultural drought that severely reduces crop yield, even though deeper soil levels may be saturated. The onset of an agricultural drought may follow a meteorological drought, depending on the prior moisture status of the surface soil layers. Precipitation deficits over a prolonged period that affect surface or subsurface water supply, thus reducing streamflow, groundwater, reservoir and lake levels, may lead to a hydrological and ground water drought, which will persist long after a meteorological drought has ended (Heim, 2002). The ground water drought, can be different from hydrological drought due to the involvement of complex hydrological processes (Mishra and Singh, 2010). Socioeconomic drought associates the supply and demand of some economic goods with certain elements of meteorological, agricultural, and hydrological droughts. The relationship between hydroclimatic variables and different types of droughts is complex and hence it is difficult to develop an accurate index to quantify and compare droughts.

Currently, there exist a number of drought indices that are used to represent different types of droughts. Some of the earlier drought indices include: Munger's Index (Munger, 1916),

* Corresponding author.
E-mail address: deepthir86@gmail.com (D. Rajsekhar).

http://dx.doi.org/10.1016/j.jhydrol.2014.11.031
0022-1694/

Blumenstock's Index (Blumenstock, 1942), and Antecedent Precipitation Index(McQuigg, 1954) which are all basically precipitation based indices. In 1965, Palmer (1965) introduced the widely popular Palmer Drought Severity Index (PDSI) which is based on precipitation and temperature as input variables in a water budget model. Despite its wide usage, it has several limitations like lack of physical meaning, slowness in detecting the onset of drought events, unclear temporal scale and problems with Thornthwaite's method used for calculation of PDSI. McKee et al. (1993) introduced another popular drought index named as Standardised Precipitation Index (SPI). SPI has several advantages like comparability among various locations, and wide range of time scales ranging from 1-month to 24-months. However, multiple SPIs with various time scales may also lead to confusion in assessment of drought condition. Similar to SPI, there are other indices like Standardized Runoff Index (SRI; Shukla and Wood, 2008) and Standardized Streamflow Index (SSFI; Modarres, 2007) which use runoff and streamflow as drought indicator variables. Other commonly used indices include Crop Moisture Index (CMI; Palmer, 1968) for agricultural drought, Vegetation Condition Index (VCI; Kogan, 1995), Climate prediction center (CPC) Soil Moisture Index (SMI; Huang et al., 1996), and Standardized Precipitation Evapotranspiration Index (SPEI; Vicente-Serrano et al., 2010).

All of these indices consider one specific physical form of drought: hydrological, meteorological, or agricultural. This might not be adequate to get a comprehensive idea of the drought condition since it is dependent on multiple variables. Hence, in general it can be concluded that the drought status indicated by one drought index might not be consistent with the findings obtained while using a different drought index.

To overcome these limitations, a group of indices that consider multiple variables to represent drought were developed. The drought monitor developed by Svoboda et al. (2002) considers an Objective Blend of Drought Indicators (OBDI) which is the linear weighted average of several drought indices. Aggregated Drought Index (ADI; Keyantash and Dracup, 2004) comprehensively considers all physical forms of drought through variables like precipitation, streamflow, evapotranspiration, reservoir storage, soil moisture content and snow water content. ADI aggregates all these variables into a single time series through principal component analysis (PCA). However, the use of PCA has several limitations like linearity assumption in data transformation, and the assumption that most information is contained in those directions where input data variance is maximum. These assumptions however need not be always met in reality. Recently, bivariate drought indices have been derived using copulas to quantify the joint behavior of drought types. Kao and Govindaraju (2010) introduced a Joint Drought Index (JDI) using copula for obtaining the joint probabilities while considering precipitation and streamflow. Hao and AghaKouchak (2013b) introduced Multivariate Standardized Drought Index (MSDI) which uses copula to form joint probabilities of precipitation and soil moisture content. The use of copula for multivariate analysis is, no doubt, highly effective. However, for higher dimensional cases (i.e., more than three variables), this method will not be a feasible choice due to the lack of flexibility in modeling the dependence structure.

Feature extraction technique is an effective approach to aggregate the various drought types into a single index. The PCA, which has been commonly used in hydrology and water resources, is a popular technique that falls under the class of linear feature extraction models. Over time, other techniques were developed, which tackled the non-linearity problem through local approaches (Roweis and Saul, 2000), neural networks (Kramer, 1991), or kernel algorithms (Scholkopf et al., 1999). The kernel based methods, like the kernel principal component analysis (KPCA) and kernel partial least squares (KPLS), have attracted a lot of attention, particularly in the last decade as an effective non-linear approach for dimensionality reduction. These methods target at finding projections that maximize the variance of input data in the feature space. However, the method assumes that the maximum information that can be obtained from the input data is oriented along the direction of maximum variance. It has been proved that entropy is a much better measure of information than variance (Dionisio et al., 2007). Entropy is related to the higher order moments of a distribution, and thus, unlike the variance, it can offer a better characterization of the input data, since it uses more information from the probability distribution (Ebrahimi et al., 1999).

The objective of this study, therefore, is to make use of a kernel entropy component analysis (KECA) for extracting a drought index named as multivariate drought index (MDI) from the set of input variables representing the various physical forms of drought. We consider the variables: precipitation (P), runoff (R), evapotranspiration (ET), and soil moisture (SM), thus accounting for all the major elements in the water balance model. The method is essentially a novel feature extraction technique that combines the concept of entropy and KPCA. The KECA or entropy PCA performs dimensionality reduction by projecting the data onto those kernel principal component axes that maximally contribute to the entropy estimate of the input dataset. These axes will not necessarily correspond to the top eigenvalues or eigenvectors of the kernel matrix, as in the case of KPCA (Jenssen, 2010). The KECA thus overcomes the disadvantages of PCA and KPCA. The advantages of KECA include: (1) It does not make the linearity assumption; (2) final multivariate index is obtained in such a way that it preserves the entropy of the input data, which means it tries to preserve the maximum amount of information of the input data; and (3) unlike KPCA, it does not make the assumption that the maximum information from the input data is oriented along the direction of maximum variance. KPCA essentially preserves only the second order statistics of data set, whereas KECA preserves the higher order statistics also through the use of entropy. Additionally, this study also explored the multiscalar nature of MDI by comparing the entropy values of different temporal scales. This would guide the user to choose the most suitable time scale required for further analysis or decision making.

The paper is organized as follows. The second section deals with the study area. Section three discusses data, its sources and the description of the model used for simulating the input variables. The methodology is described in section four, followed by results in the fifth section. The sixth section discuss the results and the conclusions drawn from the study.

2. Study area

The study area considered is the state of Texas in the USA. It is the second largest state in United States with a total land area of 261,914 square miles. Because of its size and geographical location, the state has a diverse climate ranging from arid to subtropical humid. There are five distinct climate zones in Texas, namely arid, semi-arid, continental steppe, sub-tropical semi-humid and sub-tropical humid zones. The basic climatic pattern within Texas is fairly simple: annual mean temperature increases from north to south and annual mean precipitation increases from west to east. Hot spots are found in Rio Grande and Red River Basin, whereas the mountains in west Texas experience the coolest summertime temperatures (Nielson-Gammon, 1995). The varied physiography in Texas from forests in the east and 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 arid climate zone covering higher

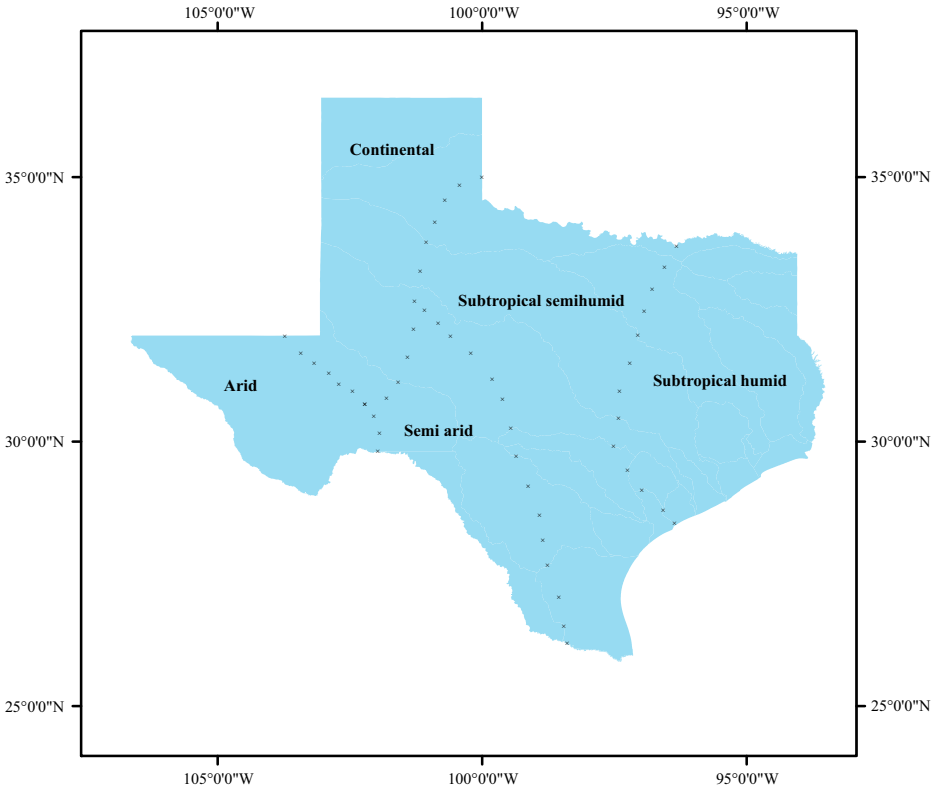

Fig. 1. Climate zones of Texas.

elevation areas, whereas most of the sub-tropical humid zone and parts of sub-tropical semi-humid zone cover low lying regions. 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. In Sabine River basin in east Texas, mean annual rainfall is nearly 60 inches and annual evaporation is less than 70 inches, whereas in Rio Grande basin in west Texas, the mean annual rainfall ranges from 8 to 20 inches and annual evaporation is as much as 105 inches. Therefore, east Texas rivers flow year around, whereas most of the west Texas streams flow only part of the year (Bureau of Economic Geology, 1996).

Texas is a highly drought prone state. Some of the biggest drought events in the history of Texas include the dust bowl drought that spanned during the time period 1933–1940. This was followed by the 1950s drought that lasted from 1950 to 1957. The 2010–2011 drought that followed also became one of the costliest natural hazards the state had ever encountered. Fig. 1 shows the five climatic regions within Texas.

3. Data

The hydroclimatic variables considered for deriving MDI include: precipitation (P), runoff (R), evapotranspiration (ET) and soil moisture (SM) for a time period of 1950–2012 on a monthly time scale. A large scale hydrologic model, called variable infiltration capacity (VIC) model, was used to generate R, ET and SM for the state of Texas (www.hydro.washington.edu). The VIC-3L model was used for simulating land–atmosphere fluxes by solving the water and energy balance at a daily or sub-daily temporal scale. The land surface is essentially divided into grids of specified resolution (default is a latitude–longitude of 1/8°). Each of these cells is simulated independent of each other. The soil moisture distribution, infiltration, drainage between soil layers, surface runoff, and subsurface runoff are all calculated for each grid at each time step. Liang et al. (1994) gives more details regarding the working of the VIC model. The input files required for running the VIC model include daily meteorological data (precipitation, wind speed, and temperature), soil and land cover data, all at the specified resolution of 1/8°. The VIC model overcomes the limitations that exist due to the lack of long-term observed databases in the case of soil moisture and evapotranspiration, and non-uniform distribution of streamflow gauging stations which are mostly concentrated in eastern Texas and poorly distributed in western Texas, thus providing fine scale data that is essential to account for the spatial heterogeneity of droughts. A brief discussion on hydroclimatic variables used in this study is now given.

3.1. Precipitation

Precipitation data was obtained from National Oceanic and Atmospheric Administration (NOAA) Cooperative Observer (Co-op) stations. The precipitation gauge data were gridded to the 1/8° resolution using the synergraphic mapping system (SYMAP) algorithm of Shepard (1984). This gridded precipitation data was then scaled to match the long-term average of the parameter-elevation regressions on independent slopes model (PRISM) precipitation climatology, which is a comprehensive dataset of 12 monthly means for the period of 1961–1990 that is statistically adjusted to capture local variations due to complex terrain. The scale factor would be the ratio of mean monthly PRISM precipitation for the period 1961–1990 to the unscaled mean monthly observed precipitation for the grid during 1961–1990. For each grid, there would be a different scaling factor for each month.

3.2. Runoff

Monthly runoff values for grids of 1/8° resolution in Texas were obtained using the VIC model. However, the VIC model does not account for interflow between grids. Because of this, to validate

the model simulation results, a routing model should be used as a post processing tool to produce streamflow at the points of interest. Lohmann et al. (1996, 1998) explains the mathematical formulation of the stand-alone routing model that was employed to transport grid cell surface runoff and base flow to the outlet of that grid cell and then onto the river system. In this routing scheme, the surface runoff simulated by VIC in each grid cell is 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, thus generating streamflow at points of interest. After routing, the simulated streamflow values will be validated against the USGS hydro climatic data network (HCDN) streamflow data. The HCDN streamflow data provides the naturalized streamflow data. Since the VIC model also simulates naturalized stream flow, the HCDN data is a better option over the original USGS gauge network for model validation.

3.3. Evapotranspiration

The VIC model considers three types of evaporation: evaporation from the canopy layer (E_c, mm) of each vegetation tile, transpiration (E_t, mm) from each of the vegetation tiles, and evaporation from the bare soil (E_1, mm) (Liang et al., 1994). The total evapotranspiration over a grid cell was computed as the sum of the above components, weighted by the respective surface cover area fractions. The model simulated evapotranspiration values were validated against the actual evapotranspiration data obtained from the Texas ET network and Texas Water Development Board (TWDB).

3.4. Soil moisture

The VIC model assumes that there is no lateral flow in the top two soil layers; therefore, the movement of moisture can be characterized by the one-dimensional Richards equation. The soil moisture percentiles were simulated for the top 40 cm of soil. The simulated soil moisture percentiles were validated using soil moisture data obtained from soil climate analysis network (SCAN) stations maintained by Natural Resources Conservation Service (NRCS), and climate reference network stations.

4. Methodology

The mathematical formulation of MDI has two steps as follows:

(1) The input hydro climatic variables used for the calculation of MDI were transformed into standard normal variates.
(2) The information theory based feature extraction technique called KECA was then utilized to extract the MDI time series that maximally preserved the entropy of the standardized input dataset.

The following sections discuss in detail the steps involved in the calculation of MDI.

4.1. Standardization of input variables

The primary step involved in the mathematical formulation of MDI is the transformation of each input variable into an index that is a standard normal variate. The procedure followed for the calculation of these standardized indices consists of the identification of a suitable probability distribution fitted to the monthly time series of the variable under consideration, followed by the construction of cumulative density function which is then transformed to standard normal distribution function. This is calculated by a numerical approximation of the normal cumulative distribution function (CDF). The approximation given by Abramovitz and Stegun (1964) was used to obtain the standard cumulative normal probability distribution function (CDF). The approximation for the CDF $\phi(x)$ for $x > 0$ is given by:

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

where $\phi(x)$ is the standard normal CDF, $b_0 = 0.2316419$, $b_1 = 0.319381530$, $b_2 = -0.356563782$, $b_3 = 1.781477937$, $b_4 = -1.821255978$, and $b_5 = 1.330274429$. Having approximated the normal CDF using Eq. (1), the respective standardized index for the time series of the given variable was obtained as the standard normal variate with zero mean and unit standard deviation.

Since both precipitation and evapotranspiration were considered as input variables, a combined standardized drought index popularly known as standardized precipitation evapotranspiration index (SPEI) developed by Vicente-Serrano et al. (2010) was used, instead of calculating separate standardized indices for precipitation and evapotranspiration. A differential timeseries $D = P - ET$ formed the basis of SPEI. The D timeseries was fitted to a three parameter log–logistic distribution to get the cumulative probabilities (Vicente-Serrano et al., 2010). These cumulative probabilities were converted to standard normal variates by following the steps outlined above in order to obtain SPEI. Likewise, the log-normal distribution was used to fit the runoff time series and obtain the CDFs which were subsequently converted to standard normal CDFs and a Standard Runoff Index (SRI) was obtained (Shukla and Wood, 2008; Rajsekhar et al., 2013). A non-parametric approach was used to obtain the empirical probabilities of soil moisture data using the Gringorten plotting position. The Gringorten formula is given by (Gringorten, 1963):

$$P_r = \frac{r - 0.44}{n + 0.12} \tag{2}$$

where i is the rank when the data is arranged in descending order, and n is the sample size. These empirical probabilities were then transformed to standard normal CDF and a standardized Soil Moisture Index (SMI) was obtained (Hao and AghaKouchak, 2013a,b).

Thus, the input dataset used for formulating MDI consisted of SPEI, SRI and SMI. Note that all of these indices have multiscalar property like SPI. This property was acquired by MDI as well. Thus, an n-month MDI was calculated by considering the n-month totals of each input drought variable.

4.2. Spectral methods for data transformation

The approach used for aggregating the input data set into MDI was through a data transformation technique that combined the concept of entropy with kernel principal component analysis. Data transformation techniques are basically used to convert high dimension data into an alternative lower dimensional representation that preserves the structure of the original data. Several data transformation methodologies have been reported in the literature. Spectral methods are the most popular technique used for this purpose, and it is based on the eigen values and eigen vectors of specially constructed data matrices. Saul et al. (2006) give a detailed review of the spectral methods for data transformation. This section discusses Principal component analysis (PCA) and its extension.

Principal component analysis (PCA) is a linear dimensionality reduction technique. PCA aims at developing a lower dimensional data representation of the original data in such a way that the

transformed data preserves the covariance structure. The input patterns, $X = x_1, \ldots x_n; x_i \epsilon R^d$ are projected onto an m-dimensional subspace that has maximum variance. The output obtained through PCA are the coordinates of the input data in this subspace, using the directions specified by the top m eigen vectors as the principal axes. The eigen values of the covariance matrix represent the variance in the eigen-directions of data space. Hence, the eigenvector corresponding to the largest eigenvalue is the direction in which the data is most stretched out. The second direction is orthogonal to it and picks the direction of the largest variance in that orthogonal subspace, and so on and so forth. Thus, the number of significant eigen values determines the dimensionality of the subspace that explains most of the original data's variance.

As an advancement from linear methods, a number of nonlinear spectral data transformation methods like kernel PCA have been proposed (Scholkopf et al., 1999, Scholkopf, 2000). Kernel PCA (KPCA) performs like traditional PCA in a so-called kernel feature space which is nonlinearly related to the input space. Suppose we are given a real-valued function $K : R^d X R^d \rightarrow R$ with the property that there exists a map $\phi : R^d \rightarrow H$ into a dot product feature space H such that for all x, $x6/\epsilon R^d$, we have $\phi(x).\phi(x') = K(x, x')$. The kernel function $K(x, x')$ can be viewed as a nonlinear similarity measure (Scholkopf and Smola, 2002). The covariance matrix in this case can be given as:

$$C = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\phi(x_i^T) \tag{3}$$

The top m eigenvectors of C are denoted as $(v_\alpha)_{\alpha=1}^m$ and their respective eigenvalues as $(\lambda_\alpha)_{\alpha=1}^m$. The kernel matrix K may be eigendecomposed as $K = EDE^T$, where D is the diagonal matrix storing all the eigenvalues $\lambda_1, \ldots, \lambda_n$ and E is a matrix with the corresponding eigenvectors $v_1, \ldots, v_n$ as columns (Williams, 2002). The lower dimensional outputs of KPCA are thus given by $\phi_{pca} = \sqrt{D_m}E_m^T = \sqrt{\lambda_\alpha}.v_\alpha^T$. D_m stores the top m eigenvalues of K, and E stores the corresponding eigenvectors as columns. Using the fact that the equivalence between PCA and another linear data transformation method called Metric multidimensional scaling (MDS; Borg and Groenen, 2005) holds for KPCA as well (Williams, 2002), the KPCA outputs can be seen as solution to a minimization problem which are analogous to the mathematical formulation for MDS. The minimization problem for KPCA outputs can be formulated as:

$$\phi_{pca} = D_m^{\frac{1}{2}}E_m^T : \min_{\lambda_1, v_1 \ldots, \lambda_n, v_n} I^T(K - K_{pca})^2 I \tag{4}$$

where $K_{pca} = E_m D_m E_m^T$ and I is an $(n \times 1)$ matrix of ones. KPCA shares all the statistical and mathematical properties of PCA with the modification that they become valid over the feature space H rather than R^d. Note, however, that the KPCA transformation is based on the selection of eigenvectors solely on the basis of size of eigenvalues, and hence it might end up choosing uninformative eigenvectors from an entropy perspective. To overcome this issue, a new data transformation method called kernel entropy component analysis has been employed in this study, which is explained below.

4.3. Kernel entropy component analysis (KECA)

Recently, it has been shown that there is a close connection between the kernel methods and information theory (Jenssen et al., 2005, 2006). This is a new spectral data transformation method and is fundamentally different from other spectral methods because the data transformation in this method is based on the Renyi entropy of the input space dataset. Jenssen (2010) shows that the Renyi entropy estimator of the input space can be expressed in terms of projections onto the principal axes which are the KPCA axes in kernel feature space. In KECA, the dimensionality reduction is brought about by projecting onto those KPCA axes that contribute to the entropy estimate of the input data set. In general, it need not correspond to the top eigenvalues and eigenvectors of the kernel matrix, as is the case with the KPCA method. Hence, KECA may produce strikingly different results compared to KPCA. The transformed data produced through KECA transformation shows a distinct angular structure, meaning that even nonlinearly related input datasets would be distributed in different angular directions with respect to the origin of the feature space, thus revealing more information about the input dataset.

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

$$H(X) = -\int_i P(x_i)log_2 P(x_i)dx \tag{5}$$

where $P(x_i)$'s are the probabilities associated with the events $X = x_i$'s. $H(X)$ is the marginal entropy of X which measures the information contained in X. Extensions to Shannon's entropy which result in alternate forms of information measures can be found in literature. The Renyi entropy is a more generalized and flexible form of Shannon entropy. A general form for the Renyi entropy can be given as:

$$H(X) = \frac{1}{(1-q)} \cdot log \int_i P(x_i)^q dx \tag{6}$$

The Renyi entropy becomes Shannon entropy as $q \rightarrow 1$. In this study, we focus on Renyi's quadratic entropy wherein $q \rightarrow 2$. This is the most heavily used form of Renyi entropy. The Renyi quadratic entropy is given as:

$$H(X) = -log \int_i P^2(x_i)dx \tag{7}$$

In order to estimate the Renyi entropy, we concentrate on the quantity $V(p) = \int_i P^2(x_i)dx$, which can alternately be formulated as expectation w.r.t $P(x)$, and can be calculated using the Parzen window. The Parzen window is a non-parametric density estimation method. Beirlant et al. (1997) introduced this non parametric plug-in entropy estimator. It is known for its consistency and efficiency, and provides a link between information theory and kernel learning. Using the Parzen window, the probability density estimation is given as:

$$\widehat{P}(x) = \frac{1}{n}\sum_{x_i \epsilon R_d} K_\sigma(x - x_i) \tag{8}$$

where $K_\sigma(x, x_i)$ is the Parzen window or kernel centered at x_i, and σ is the kernel size. K_σ is a Mercer kernel which is continuous, symmetric and positive semi definite. $V(p)$ can then be invoked using the sample mean approximation of expectation operator as (Jenssen, 2010):

$$\widehat{V}(p) = \frac{1}{n}\sum_{x_i \epsilon R^d}\widehat{P}(x_i) = \frac{1}{n}\sum_{x_i \epsilon R^d}\frac{1}{n}\sum_{x_i \epsilon R^d} K_\sigma(x - x_i) = \frac{1}{n^2} I^T K I \tag{9}$$

where I is an $(n \times 1)$ matrix of ones, and K is the kernel matrix. The kernel matrix K can be eigendecomposed and Eq. (9) can thus be rewritten as (Jenssen, 2010):

$$\widehat{V}(p) = \frac{1}{n^2}\sum_{i=1}^{n}\left(\sqrt{\lambda_i}v_i^T I\right)^2 \tag{10}$$

The ψ term that denotes $(\sqrt{\lambda_i}v_i^T I)^2$ contributes to the total entropy of input data. Certain eigenvectors contribute more towards the entropy than others. Eq. (10) reveals that the Renyi entropy estimator is composed of projections onto all the KPCA axes, wherein the

Table 1
Details of validation stations and time periods.

Variable	Climate zone	Data source	Station name	Latitude	Longitude	Validation period
Streamflow (CFS)	Continental	HCDN	Spring creek near Spring	35.47	−101.88	1965–1988
Streamflow (CFS)	Arid	HCDN	Frio river near Derby	31.44	−103.47	1965–1988
Streamflow (CFS)	Semi-arid	HCDN	Nueces river below Uvalde	28.50	−99.68	1965–1988
Streamflow (CFS)	Semi-humid	HCDN	Mill creek near Bellville	32.63	−101.29	1964–1974
Streamflow (CFS)	Humid	HCDN	Big cow creek near Newton	32.76	−95.46	1965–1988
Evapotranspiration (mm)	Continental	ET network, TWDB	Lubbock	33.56	−101.88	1985–2012
Evapotranspiration (mm)	Arid	ET network, TWDB	Pebble Hills park	31.78	−106.32	1985–2012
Evapotranspiration (mm)	Semi-arid	ET network, TWDB	Uvalde	29.21	−99.78	1985–2012
Evapotranspiration (mm)	Semi-humid	ET network, TWDB	McKinney	33.19	−96.63	1985–2012
Evapotranspiration (mm)	Humid	ET network, TWDB	Overton	32.27	−94.97	1985–2012
Soil moisture (percentile)	Continental	SCAN	Lehman	32.63	−102.45	2005–2012
Soil moisture (percentile)	Arid	Climate reference network	Monahans	31.58	−102.89	2010–2012
Soil moisture (percentile)	Semi-arid	Climate reference network	Edinburgh	26.30	−98.16	2010–2012
Soil moisture (percentile)	Semi-humid	Climate reference network	Austin	30.25	−97.75	2010–2012
Soil moisture (percentile)	Humid	Climate reference network	Palestine	31.75	−95.64	2009–2012

projection onto the ith principal axis is given by $(\sqrt{\lambda_i}v_i^T)$. Only a principal axis with $\lambda_i \neq 0; v_i^T I \neq 0$ contributes to the entropy estimate. Hence, a large eigenvalue λ_i simply does not guarantee that the principal axis contributes to the entropy estimate at all. The KECA transformation for an n dimensional data into a k dimensional subset is done by projecting the feature space ϕ onto a subspace ϕ_k spanned by the k kernel PCA axes that contribute most to the entropy estimate of the input data. Mathematically, this transformation can be denoted by:

$$\phi_{eca} = \phi_k^T \phi = D_k^{\frac{1}{2}} E_k^T \tag{11}$$

where D_k is the diagonal matrix containing the eigenvalues $\lambda_1, \lambda_2, \ldots \lambda_k$ that contribute the most towards the entropy of the input dataset, and E_k contains the corresponding eigenvectors $v_1, v_2, \ldots v_k$ as columns. Hence, analogous to Eq. (4), the KECA outputs could be formulated as the solution to a minimization problem:

$$\phi_{eca} = D_k^{\frac{1}{2}} E_k^T : \min_{\lambda_1, v_1 \ldots, \lambda_n, v_n} \frac{1}{n^2} I^T (K - K_{eca}) I \tag{12}$$

where $K_{eca} = E_k D_k E_k^T$. The entropy estimate of the subspace ϕ_{eca} is given as:

$$V_k(\hat{p}) = \frac{1}{n^2} I^T K_{eca} I \tag{13}$$

5. Results and discussion

Since model simulated hydro climatic variables were used for the formulation of MDI, it was essential to validate the simulations before using it for further analysis. Further, the performance of the formulated drought index was verified with respect to existing indices in the literature. The following sections discuss the results of model validation, performance of MDI, and choice of a suitable time scale for the proposed drought index.

5.1. Calibration and validation of VIC model

Since the VIC model involves a number of parameters, calibration of the same can become quite tedious. Liang et al. (1994) suggested a set of recommended parameters and the plausible range of values for each of them that can be used for calibrating the model. The VIC model calibration was performed using a random auto start simplex method program. The magnitudes of differences between simulated and observed discharge records were denoted by computing the Nash–Sutcliffe R^2 values, where 1 represents a perfect match. To explain the procedure simply, if we have n parameters for calibration, the simplex method tries to corral the minimum within a geometric shape with $n + 1$ apexes. Once a minimum has been obtained, the algorithm begins to minimize the volume of the simplex, until all of its apexes are within a specified tolerance of each other. The simplex method was applied using random autostart populations of 75–100 parameter sets. The entire cycle was repeated 5–10 times for each sub-basin. As regards the calibration of the routing model, the suggested parameters for adjustment included velocity and diffusivity. If only monthly stream flows are required, diffusivity and velocity values of 800 m^2/s and 1.5 m/s are deemed acceptable (Lohmann et al., 1996, 1998).

Three of the four variables required for the calculation of MDI was obtained through the VIC model. In the case of stream flow, the routing model was used to route the grid-cell runoff to the selected station locations. Results from the routing model were aggregated to a monthly scale and compared with the observed gauge data. The model directly simulates evapotranspiration which was compared against the observed ET values. The soil moisture averaged for each month at 40 cm soil depth was compared with the observed soil moisture obtained from SCAN stations and climate reference network. The two performance criteria selected were correlation coefficient and the Nash–Sutcliffe (N–S) efficiency. N–S efficiency (NSE) is given by:

$$NSE = 1.0 - \frac{\sum_{i=1}^{M} (O_i - S_i)^2}{\sum_{i=1}^{M} (O_i - \overline{O})^2} \tag{14}$$

where M is the length of time series, S_i is the simulated value for the ith month, O_i is the observed value for the ith month, and $\overline{O}$ is the mean monthly observed value, respectively.

It should be noted, however, that the N–S efficiency has a disadvantage in that the differences between observed and predicted values are calculated as squared values. This leads to overestimation of peak values and underestimation of trough values. In order to reduce the sensitivity of the Nash–Sutcliffe efficiency to extreme values, the logarithmic transforms of observed and simulated values can be used to calculate the N–S value (Krause et al., 2005). Through the logarithmic transformation the peaks are flattened and the low flows are kept more or less at the same level.

Table 1 gives the locations of validation stations for input variables, and the time period considered for validation. The chosen validation time period was dependent on the availability of observed data, which is sparse particularly in the case of soil moisture. Figs. 2a–c show the time series comparison between observed and simulated values of streamflow, evapotranspiration and soil moisture, respectively, at the selected locations in various climate

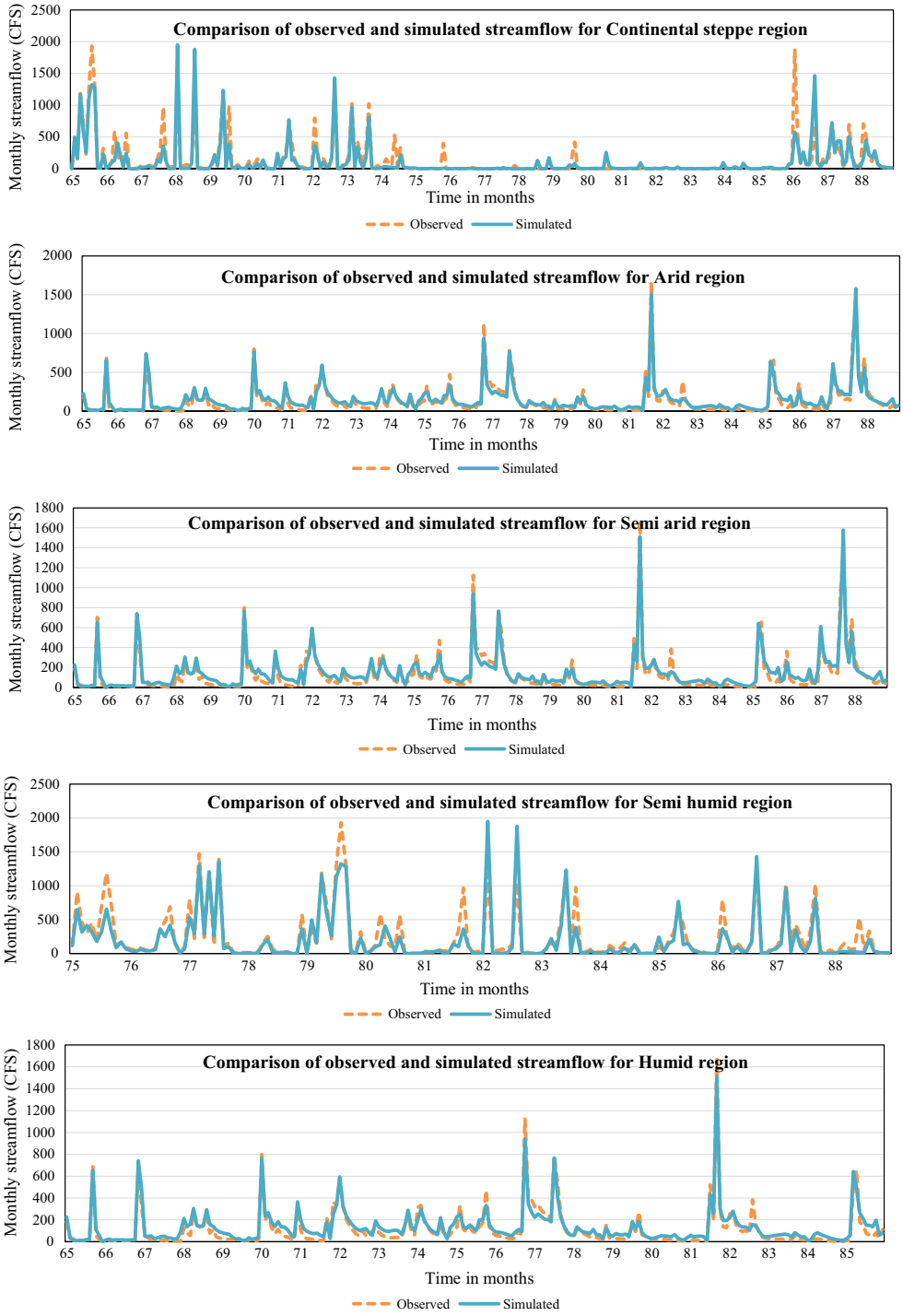

Fig. 2a. Comparison of simulated and observed streamflow for selected stations.

zones of Texas. Table 2 gives the values of performance statistics at each location considered for validation.

It can be seen from Table 2 that the correlation coefficients for streamflow validation ranges from 0.80 to 0.96, which means the model is capable of explaining 64–92% of variability in the observed data. The N–S efficiency values range from 0.54 to 0.81. Since an NSE value of 1 corresponds to a perfect match and an NSE value of 0 corresponds to the situation wherein simulated values match the mean of the observed values, a value of 0.5 may be considered to represent a mediocre model performance. Hence, from the values obtained for the model at all the 5 stations, it can be seen that the model performance is satisfactory.

In the case of evapotranspiration, the correlation coefficients fall within the range of 0.71–0.92, which means the model is explaining 50–85% of variability in the observed data. The N–S efficiency values for ET lie within the range of 0.64–0.79. Although the model replicates ET values well within acceptable limits, it seems to overpredict the values slightly, when it comes to humid climate zone, which in reality experiences the minimum evapotranspiration in Texas. In the case of soil moisture, the correlation coefficient

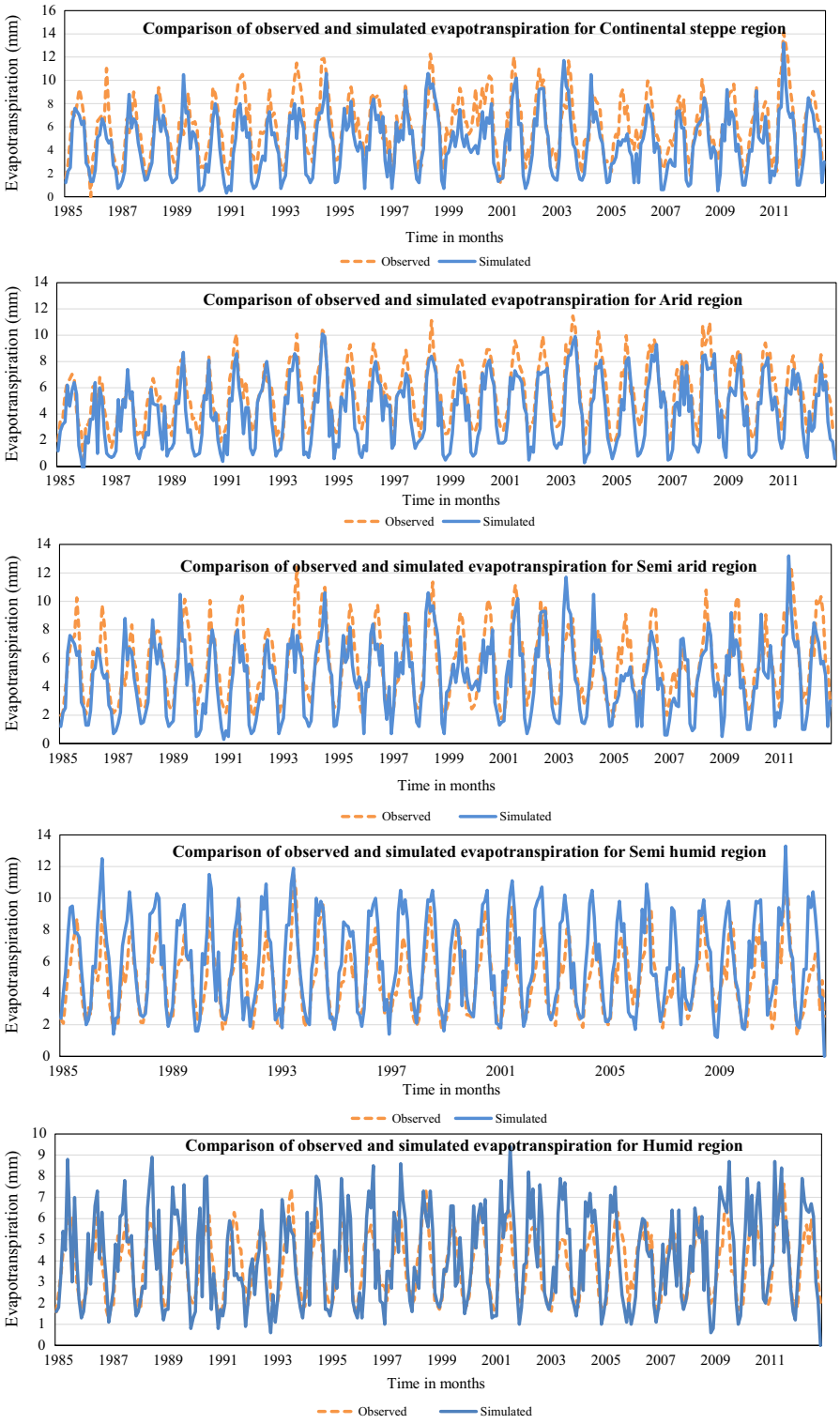

Fig. 2b. Comparison of simulated and observed evapotranspiration for selected stations.

values fall within the range 0.76–0.96, and thus the model explains 58–92% of variability in the observed data. The N–S efficiency values for soil moisture lie within the range of 0.58–0.87.

5.2. Comparison of MDI with other drought indices

In order to understand the performance of MDI in quantifying drought events, we compared it with other existing and established univariate and multivariate drought indices found in the literature. As a preliminary step for assessing the performance of MDI, it was compared with the Palmer drought severity index (PDSI). PDSI was chosen for two reasons: (1) It is widely used in the United States, and (2) it is formulated on the basis of the physical constituents of water balance. MDI also considers precipitation, runoff, evapotranspiration and soil moisture, which form the major components of the water balance. The PDSI values used

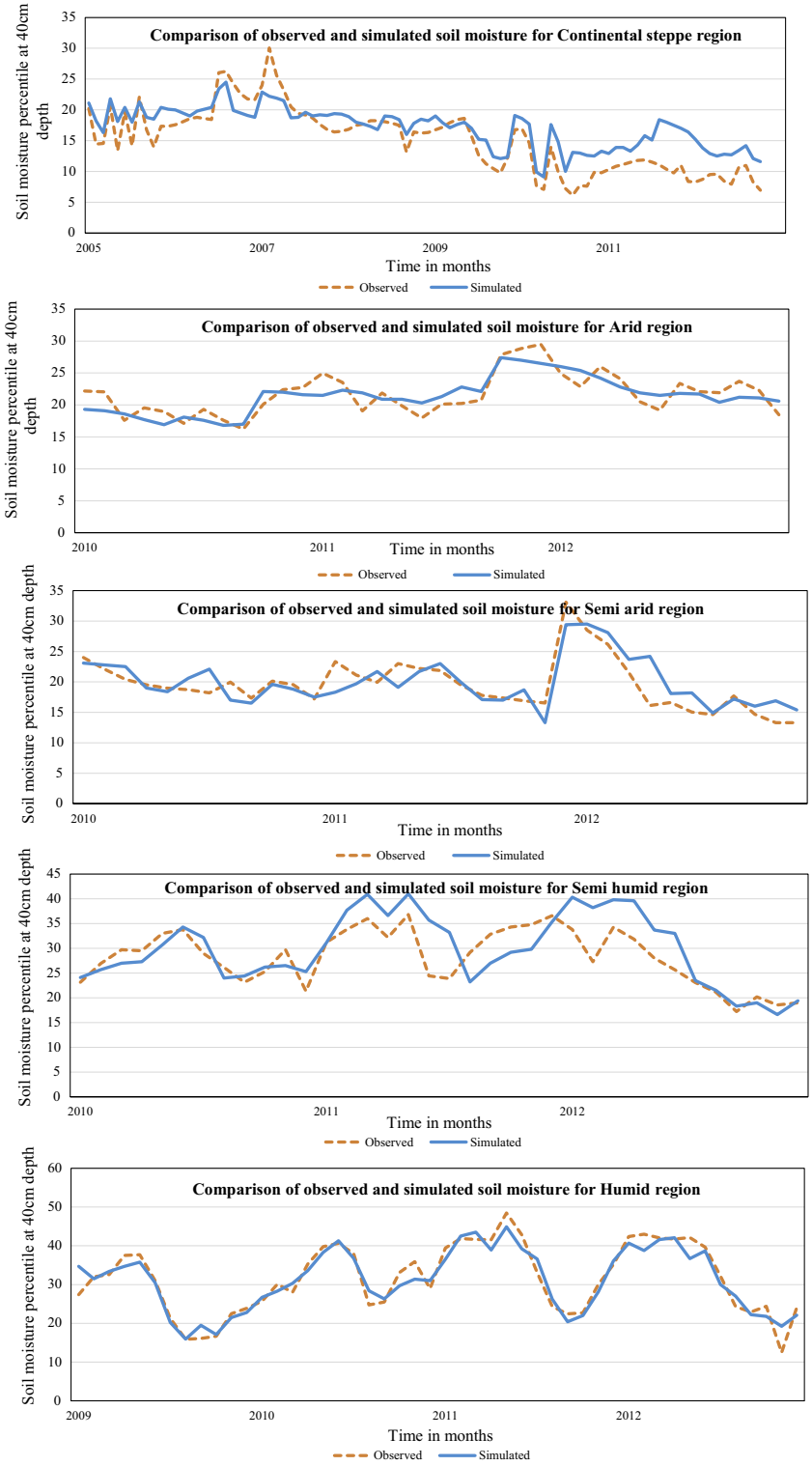

Fig. 2c. Comparison of simulated and observed soil moisture for selected stations.

for comparison were obtained from National Climate Data Center (NCDC).

The drought classification for MDI values is given in Table 3. MDI followed the same drought classification as that of its constituent indices like SPI, and SRI. Comparison of the MDI time series against PDSI for different climate zones of Texas are shown in Figs. 3a–e. Results for randomly chosen grids within each climate regions were used for visualization. The locations of the chosen grids are: (−104.75°E, 30.95°N) in arid region; (−101.37°E, 35.18°N) in continental region; (−99.93°E, 28.85°N) in semi-arid region; (−98.95°E, 32.19°N) in semi-humid region; and (−95.90°E, 30.10°N) in humid region. The portions of time series which correspond to two major Texas drought periods: 1950–1957 and 2010–2011 have been enlarged for better visualization. However, although a perfect correlation is not expected between MDI and PDSI, it is natural that they both might follow a general

Table 2
Goodness of fit test values of model validation at the selected stations.

Variable	Climate region	Correlation coefficient	N–S efficiency
Streamflow (CFS)	Continental	0.80	0.54
Streamflow (CFS)	Arid	0.93	0.77
Streamflow (CFS)	Semi-arid	0.96	0.81
Streamflow (CFS)	Semi-humid	0.85	0.75
Streamflow (CFS)	Humid	0.88	0.74
Evapotranspiration (mm)	Continental	0.85	0.71
Evapotranspiration (mm)	Arid	0.92	0.79
Evapotranspiration (mm)	Semi-arid	0.82	0.65
Evapotranspiration (mm)	Semi-humid	0.82	0.68
Evapotranspiration (mm)	Humid	0.77	0.64
Soil moisture (percentile)	Continental	0.89	0.58
Soil moisture (percentile)	Arid	0.82	0.60
Soil moisture (percentile)	Semi-arid	0.81	0.80
Soil moisture (percentile)	Semi-humid	0.76	0.72
Soil moisture (percentile)	Humid	0.96	0.87

Table 3
MDI drought classification.

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

behavioral pattern. Hence, it makes sense to analyze the monotonic relationship between MDI and PDSI. For this purpose, Spearman's rank correlation was used. Spearman's rank correlation (ρ) is a nonparametric statistical dependence measure that has many advantages over the Pearson correlation coefficient, since it does not depend upon the distribution of data and is specifically designed to study monotonic relationship between variables. Table 4 shows Spearman's ρ between PDSI and MDI in all five climate zones of Texas. It can be seen that in all climate zones, a positive monotonic relationship is seen between PDSI and MDI, and the maximum value reaches 0.71. It can be seen from Table 4 that the correlation between PDSI and MDI is strongest for the continental steppe climate zone which is characterized by low precipitation and mild winters. Following that, a relatively better correlation between the two indices can be seen for the humid region which is characterized by relatively higher precipitation. The correlation is weak for arid region and transition zones like semi-arid and semi-humid zones. Arid and semi-arid regions, in particular, are characterized by higher rates of evapotranspiration and lower precipitation. It can be seen that MDI and PDSI quantify droughts differently, particularly for transition climate zones and regions with higher rates of evapotranspiration.

Fig. 4(a)–(e) shows the comparison of annual means and variances in PDSI and MDI for different climate regions. From Fig. 4(a)–(e), it can be seen that higher drought variations are prevalent in continental steppe and arid climate zones. The annual average values of MDI and PDSI are in agreement in all climate zones other than transition zones like semi arid and semi humid regions. Fig. 4f shows the box plots of PDSI and MDI time series for different climate regions. It can be seen from Fig. 4f that the PDSI time series showed wet outliers in continental steppe and semi arid climate zones. Continental steppe region showed the lowest median out of all zones, thus indicating comparatively higher dryness. MDI and PDSI box plots show more or less similar central values, except for continental steppe wherein MDI show a lower median value than PDSI. A larger drought variability was observed in continental steppe and arid climate zones. MDI and PDSI time series for all the climate zones were more or less symmetric. However, MDI time series showed a slight left skewness in continental steppe region, and a slight right skewness in arid region.

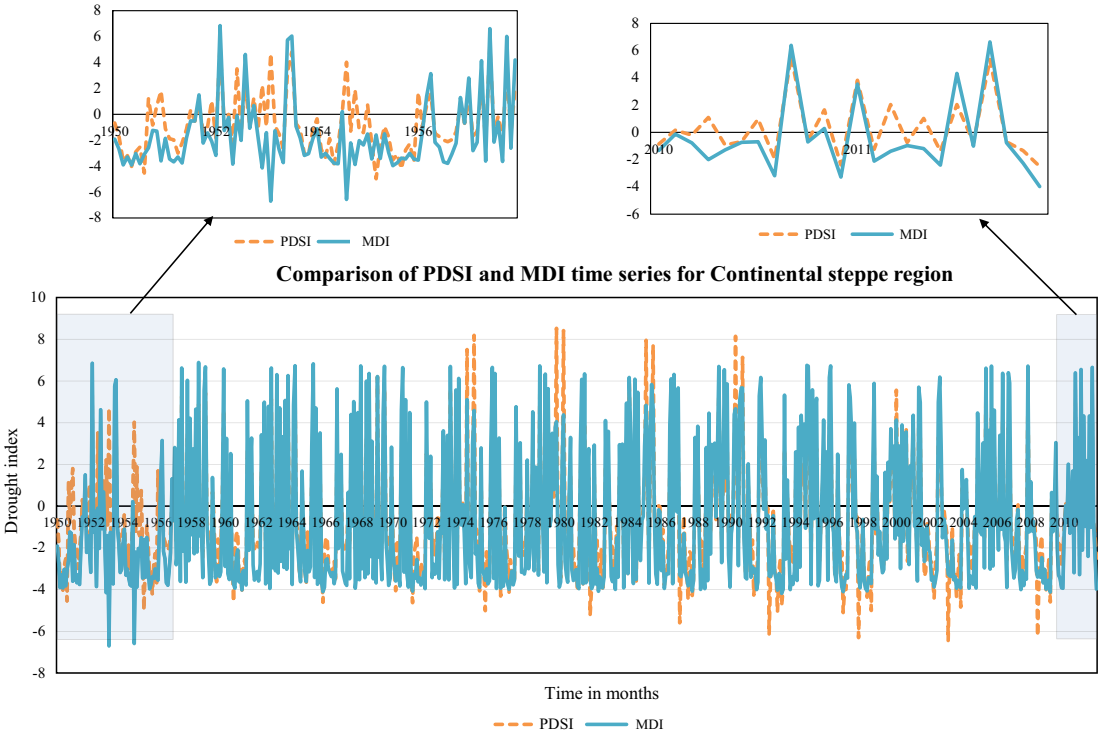

Fig. 3a. PDSI and MDI time series for continental steppe region during 1950–2012.

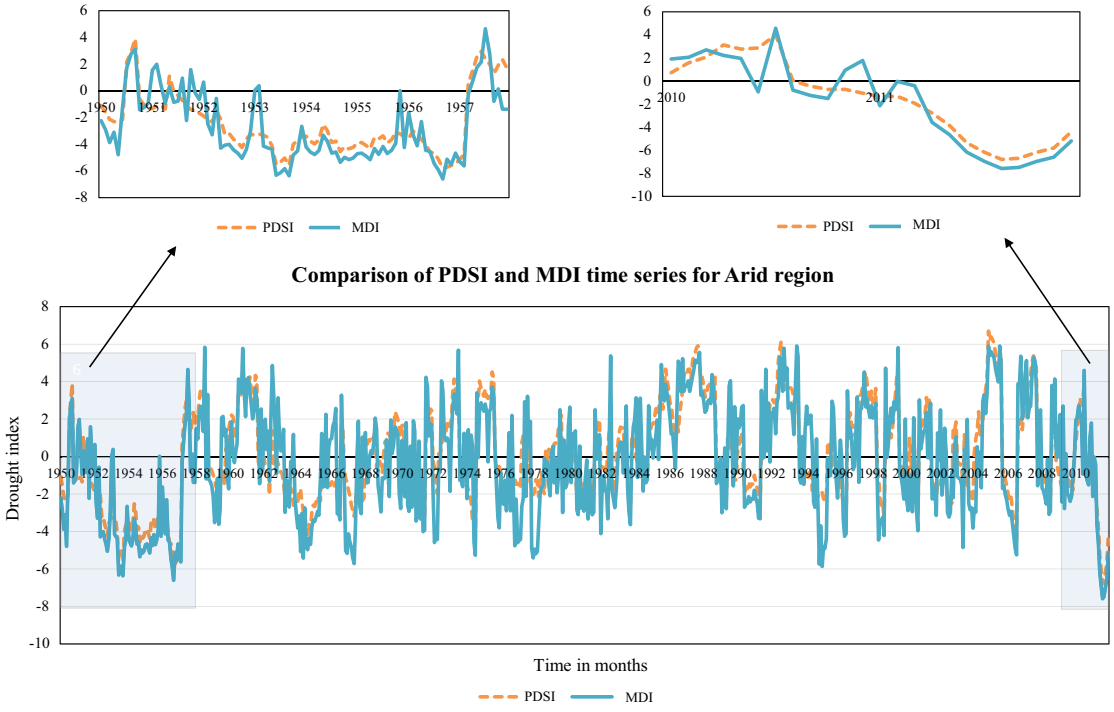

Fig. 3b. PDSI and MDI time series for arid region during 1950–2012.

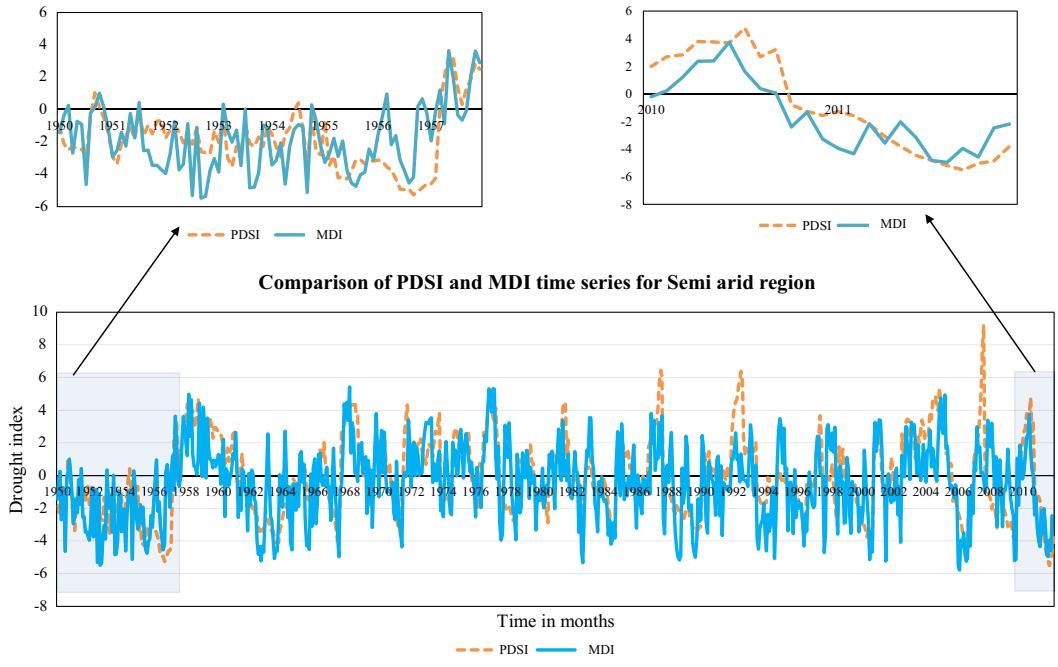

Fig. 3c. PDSI and MDI time series for semi arid region during 1950–2012.

To have a better understanding of how MDI captures the various physical forms of drought, a random location in Texas was chosen, and MDI for a short time period was compared with SPI, SRI, SMI, PDSI, MSDI (Hao and AghaKouchak, 2013a,b) and JDI (Kao and Govindaraju, 2010). The location chosen had the latitude: 30.95° and longitude: −104.75°, and falls in the arid climate zone. Time period windows considered for comparison were: 1956–57 and 2010–11, each of which falls under two major drought periods experienced in Texas. Note that 1956–57 represents the ending period for the 1950–57 drought and 2010–11 represents the beginning of another severe drought. By choosing these two windows, it was possible to see how well the indices captured the beginning and end of a drought event. Fig. 5a and b shows the time series comparison between MDI, SPI, SRI and SMI, during 1956–57 and 2010–11 time periods, respectively. Fig. 5c and d shows the time series comparison between MDI and multivariate drought indices like PDSI, MSDI and JDI, during 1956–57 and 2010–11 time periods, respectively.

It is worth mentioning here that SPI represents the meteorological drought, which may start and end rapidly; SRI represents

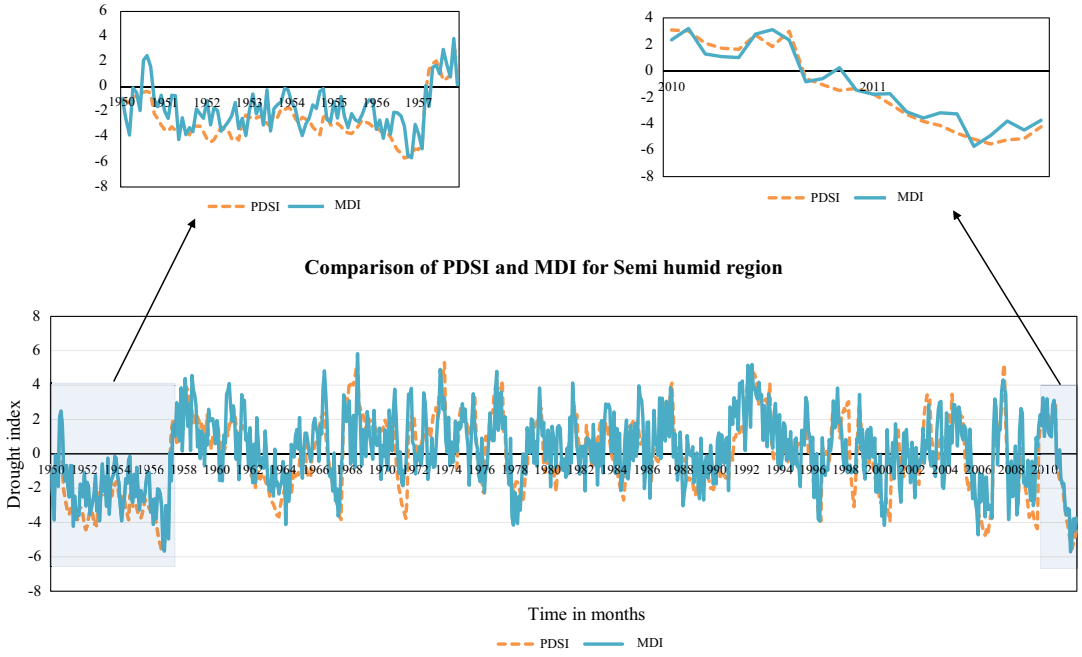

Fig. 3d. PDSI and MDI time series for semi humid region during 1950–2012.

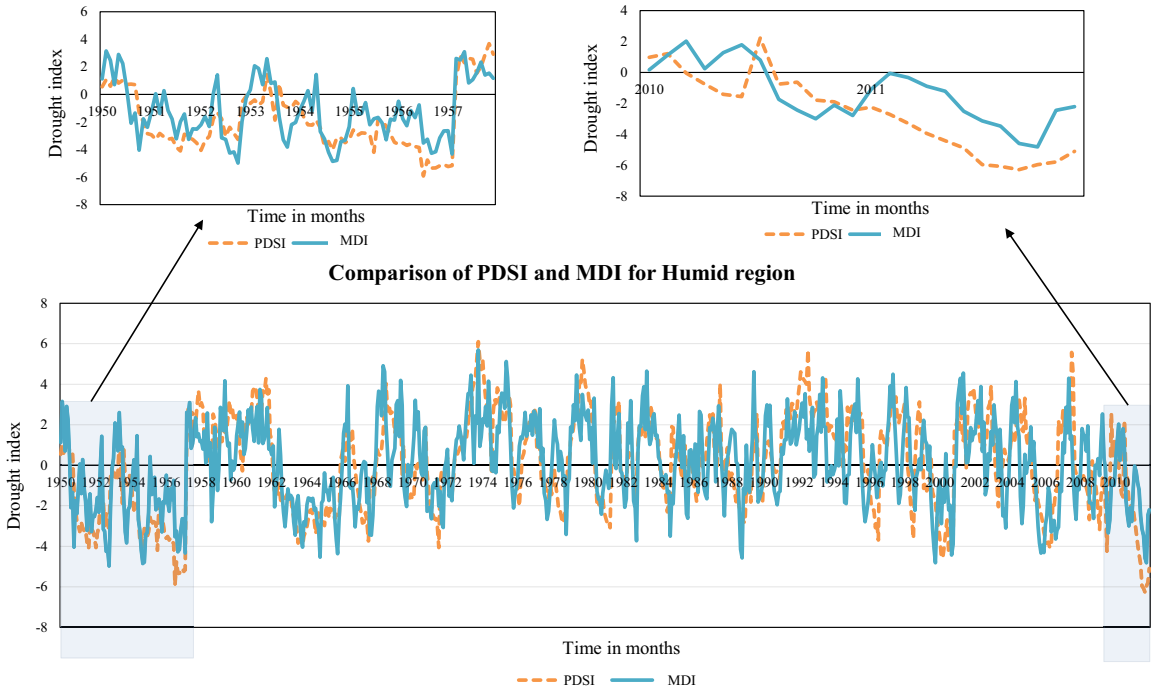

Fig. 3e. PDSI and MDI time series for humid region during 1950–2012.

hydrological drought which develops slowly and may last for a longer time, since it recovers slowly; SMI represents agricultural drought, the onset of which is determined by other factors like soil type, and temperature. It is possible that a region might be experiencing only a single type of drought at times, but multiple drought types affecting a region is quite common. MSDI indicates the combined effect of SPI and SMI, whereas JDI indicates the combined influence of SPI and SRI. Hence, the drought onset and ending as well as the magnitude predicted by each of these indices would

Table 4
Spearman's rank correlation between PDSI and MDI.

Climate zone	Spearman's rank correlation
Continental	0.71
Arid	0.57
Semi-arid	0.51
Semi-humid	0.54
Humid	0.61

be different. What we need to find out, however, is whether MDI which incorporates multiple drought forms can effectively predict the onset, termination and magnitude of drought events.

From Fig. 5a, it can be seen that if we use SPI for drought quantification, it had predicted a no-drought condition during 1956–57. This could be because meteorological droughts tend to last for a shorter time and hence the termination of the same might happen

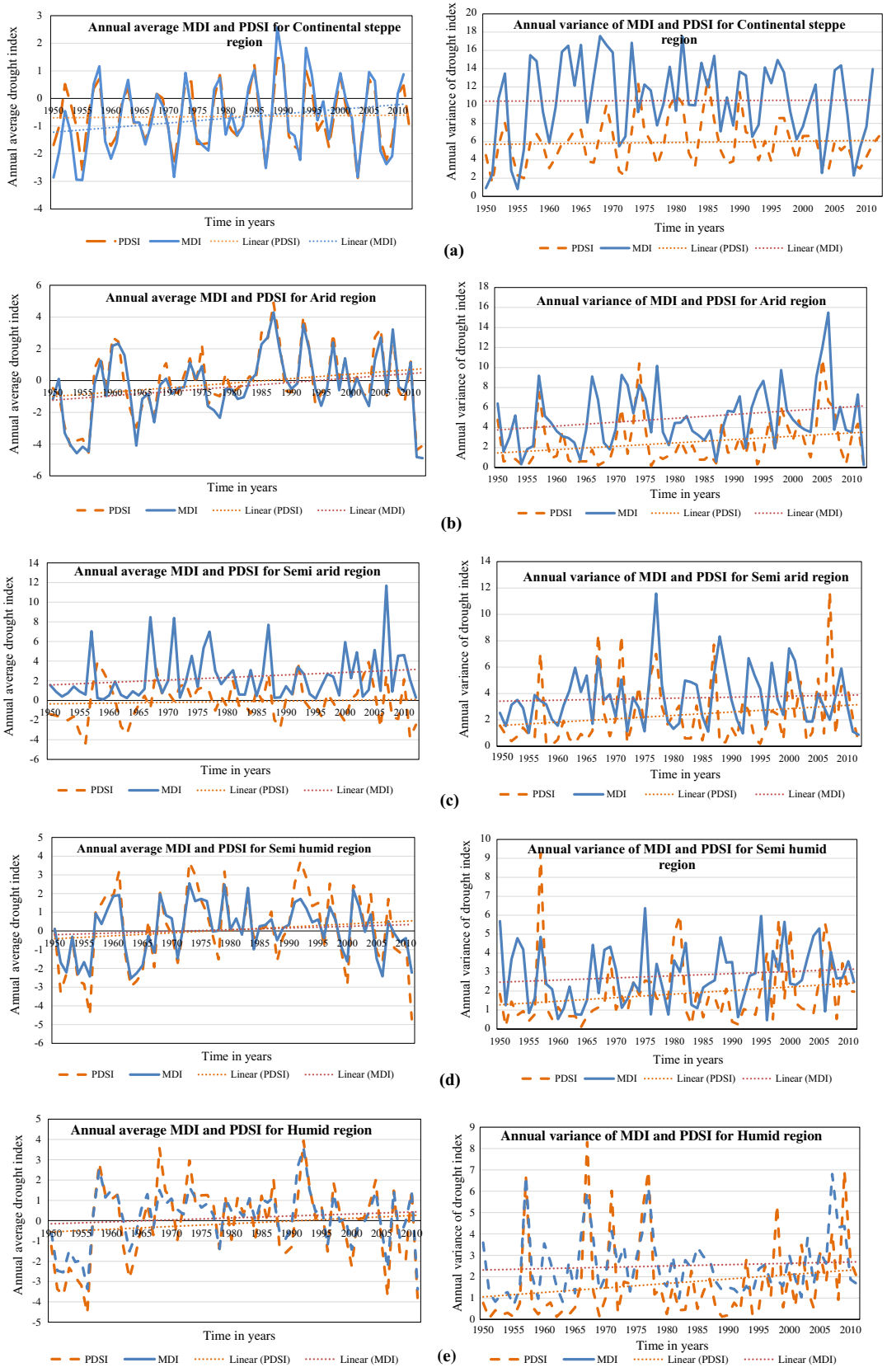

Fig. 4. Comparison of MDI and PDSI annual average mean and variance for (a) continental steppe, (b) arid, (c) semi arid, (d) semi humid, and (e) humid regions.

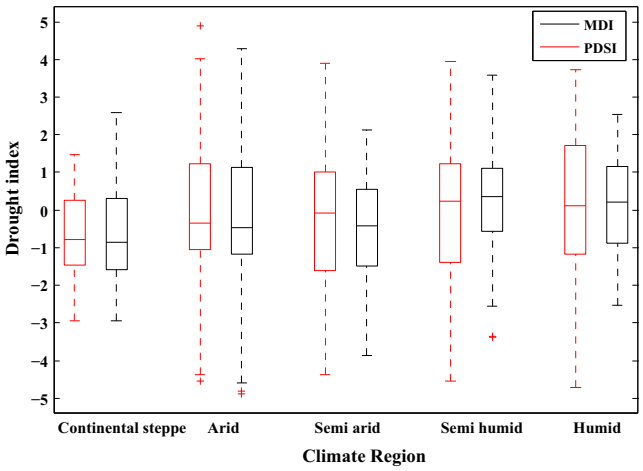

Fig. 4f. Box plots of PDSI and MDI for different climate regions.

before that of other forms of drought. If we use SRI for quantifying drought, it can be seen that it predicted drought conditions with a few fluctuations until December 1956. Thereafter, it gradually changed to a no-drought condition. This could be because hydrological drought takes a longer time than a meteorological or agricultural drought to recover. In the case of SMI, it can be seen that a drought condition was indicated mainly during February–April 1957. As explained earlier, agricultural drought which reflects the soil moisture might be influenced by other external factors like soil type and temperature. The reason for an agricultural drought could be due to a very hot summer even when there was rainfall (since there was no meteorological drought indicated during this period) which might have resulted in rapid loss of soil moisture. Since the selected location falls under arid region, higher evapotranspiration rates are expected. Thus, it can be seen from the above that none of these drought indices predict the fluctuations in the drought in a similar manner.

Hence, if we rely on one of these indices alone we might end up predicting an actual drought condition as a no-drought condition. When we used MDI, it was seen that it showed the presence of drought throughout January 1956 till July 1957 with some minor fluctuations in between. After July 1957, the drought condition was shown to improve until it started showing the start of another drought event starting on November 1957. Hence, MDI captured

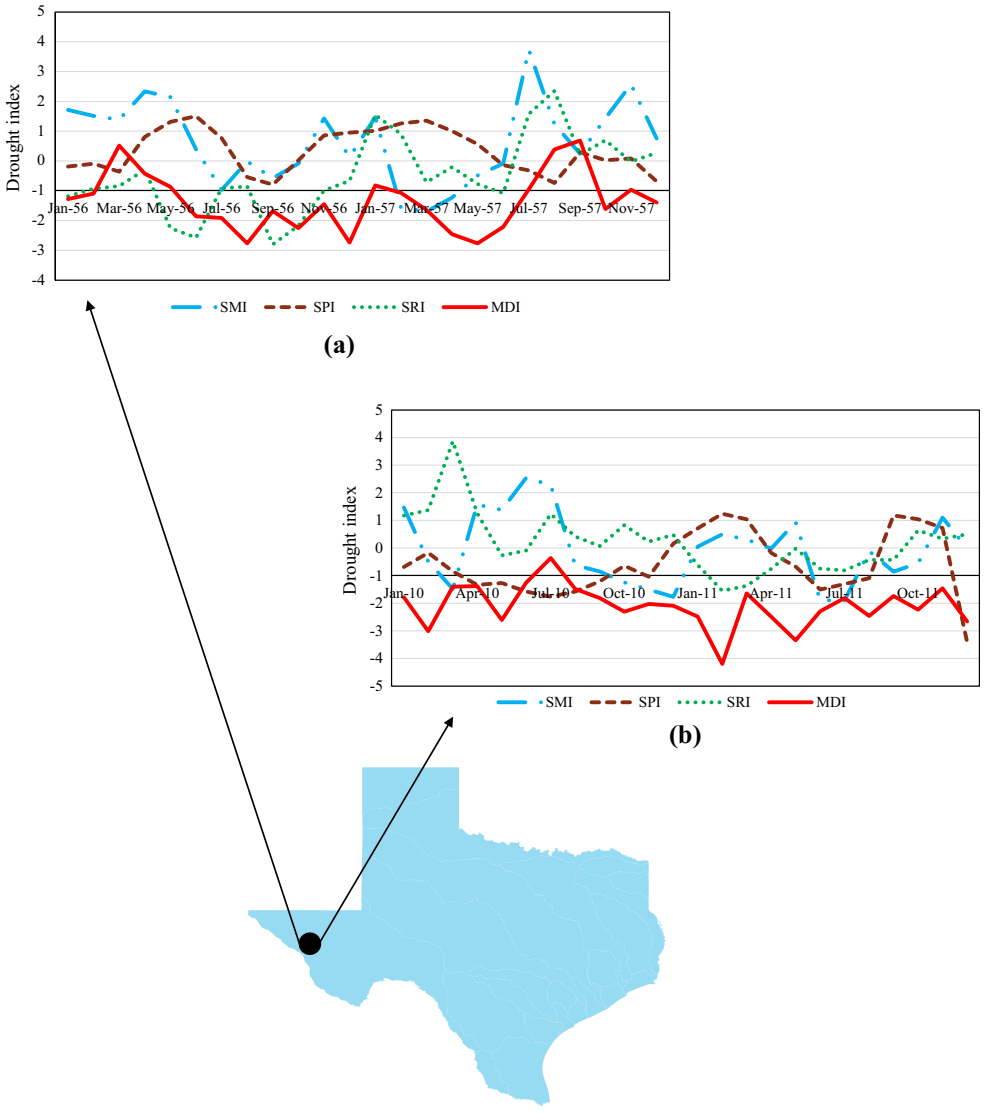

Fig. 5a and b. Comparison of MDI with univariate drought indices during (a) 1956–57, (b) 2010–11.

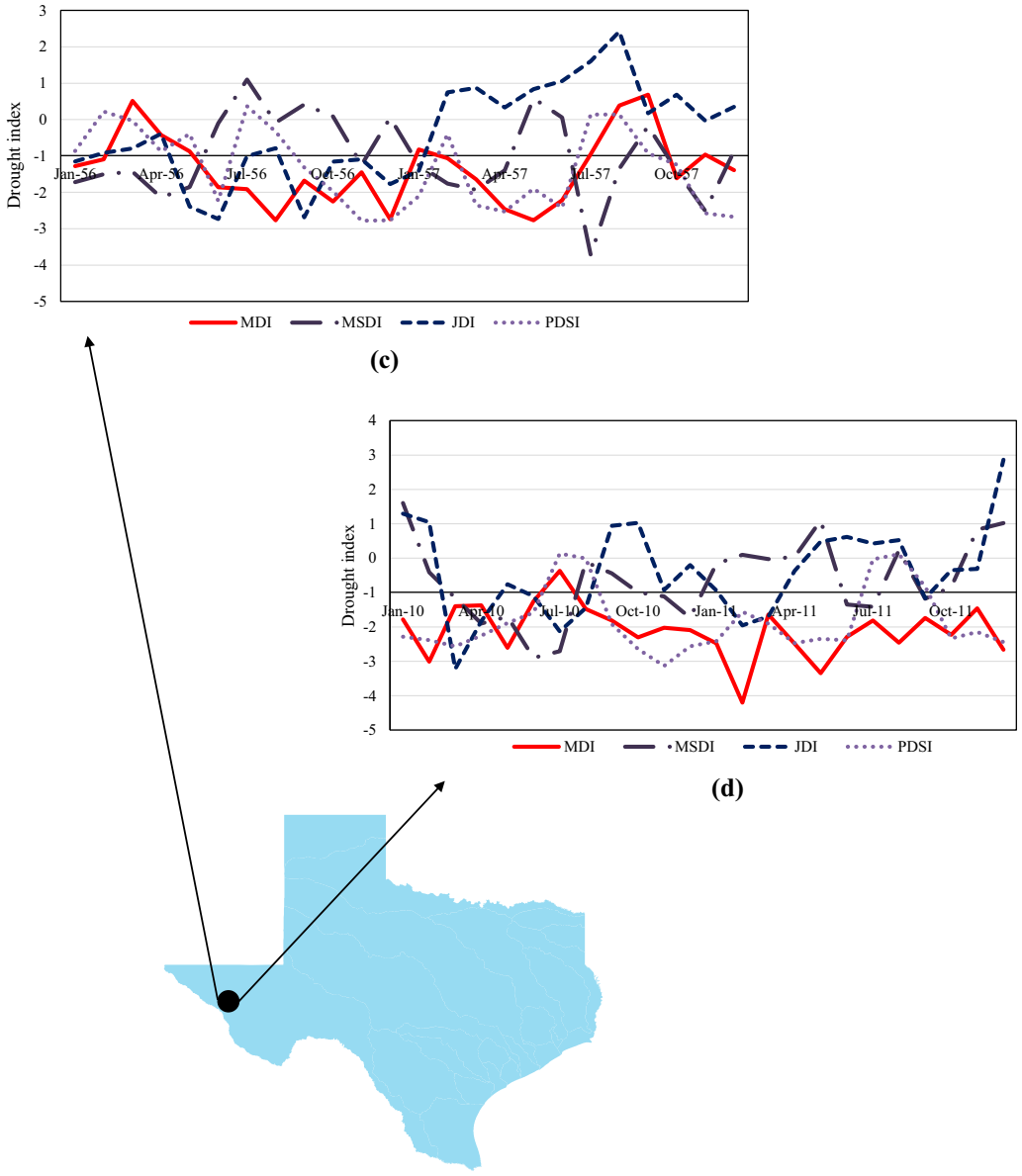

Fig. 5c and d. Comparison of MDI with multivariate drought indices during (c) 1956–57 (d) 2010–11.

the actual drought condition more effectively by combining the effects of SPI, SRI and SMI.

Now let us consider the time period 2010–2011. This indicates the beginning of a major drought experienced in Texas. From Fig. 5b, it can be seen that SPI predicted the onset of drought much ahead of SMI and SRI. This is because meteorological droughts set in faster than other physical drought forms. SRI was the last index to recognize the onset of drought. This could be because hydrological droughts typically develop gradually. MDI predicted the onset of drought at January 2010 itself. It also predicted a prevailing drought condition during most of 2010–11.

From Fig. 5c it can be seen that MSDI and JDI predicted the presence of drought as opposed to SPI which indicated a no-drought condition. This is because in addition to SPI, MSDI incorporated the effect of soil moisture drought, and JDI considered the effect of hydrological drought. MSDI and JDI showed a faster recovery from drought condition when compared to MDI. Thus, MDI showed better drought persistence property compared to bivariate drought indices. From Fig. 5d it can be seen that JDI and MSDI detected the onset of drought better than SMI or SRI due to the incorporation of SPI in their respective formulations. Both MDI and PDSI detected the drought onset earlier than MSDI and JDI.

From the two scenarios, we can see that SPI is usually good at predicting the onset of drought, whereas SRI is better at capturing the termination of drought. MDI captures both, thus making it superior to univariate drought indices. Both MSDI and JDI performed better than univariate drought indices. MDI captured the drought persistence and showed early drought detection property. PDSI followed a similar pattern to MDI, but MDI overcomes all the disadvantages that PDSI has, with its solid information theory background and multi scalar nature being the leading advantages. MDI was also better at capturing the finer fluctuations in the drought condition compared to PDSI.

Having seen that different drought indices predicted the onset and termination of drought events differently, it is expected that the values of drought properties like severity and duration quantified by these indices would also be different. We mapped the drought severity and duration patterns predicted by SPI, SRI, SMI, MDI and PDSI for the two major drought events: 1950–57 and 2010–11. The theory of runs (Yevjevich et al. (1967)) was used to

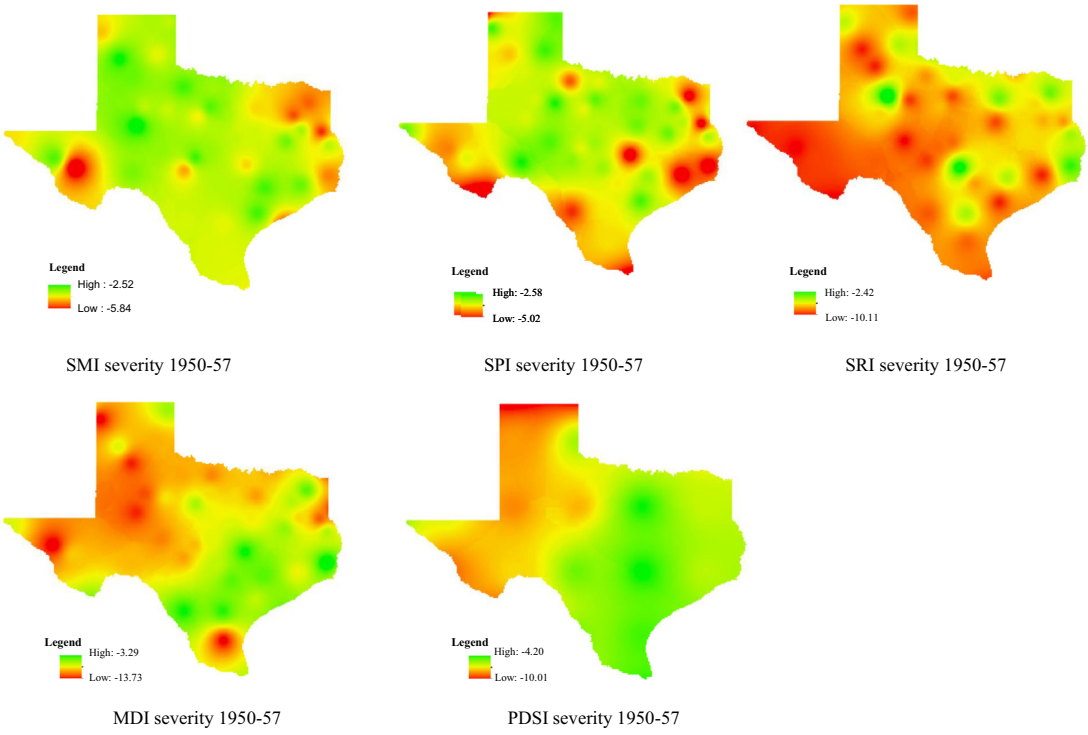

Fig. 6a. Drought severity for Texas during 1950–57 drought using various indices.

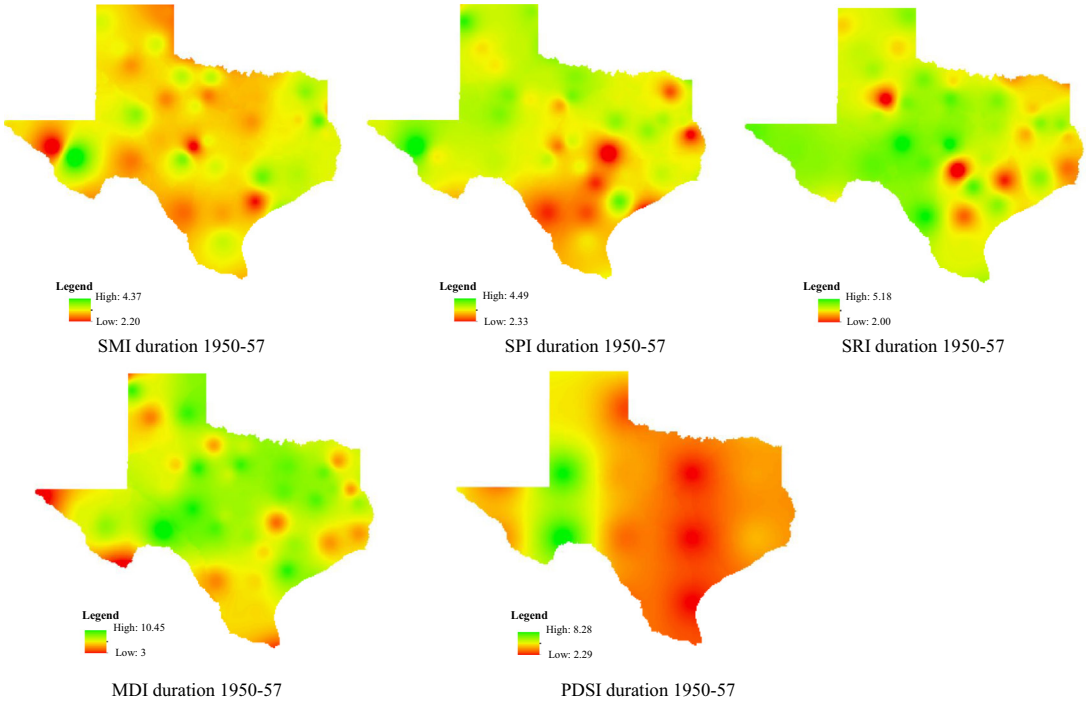

Fig. 6b. Drought duration for Texas during 1950–57 drought using various indices.

derive the drought properties of severity and duration. A threshold of −0.99 was chosen so that a value below that indicated a drought event.

Figs. 6a and 6b shows the maps of average drought severity and durations calculated using different indices during the time period 1950–57, and Fig. 7a and 7b shows the maps of average drought severity and durations calculated using different indices for the time period 2010–11. Table 5 shows the summary of drought properties for the time periods 1950–57 and 2010–11, while different indices were used for analysis. It can be seen from Table 5 that for the drought period 1950–57, MDI predicted the most severe drought, having a magnitude of −13.73. The drought duration values were also higher for predictions made using MDI, with a maximum average duration predicted as 10.45 months, closely

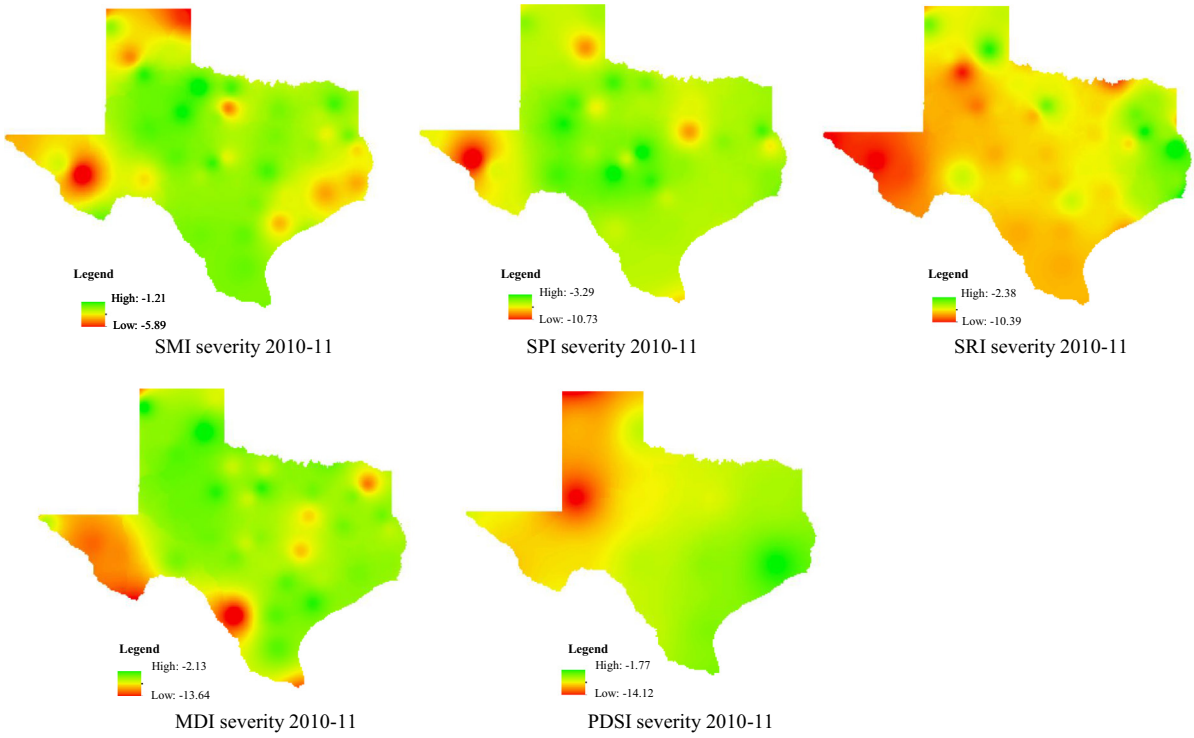

Fig. 7a. Drought severity for Texas during 2010–11 drought using various indices.

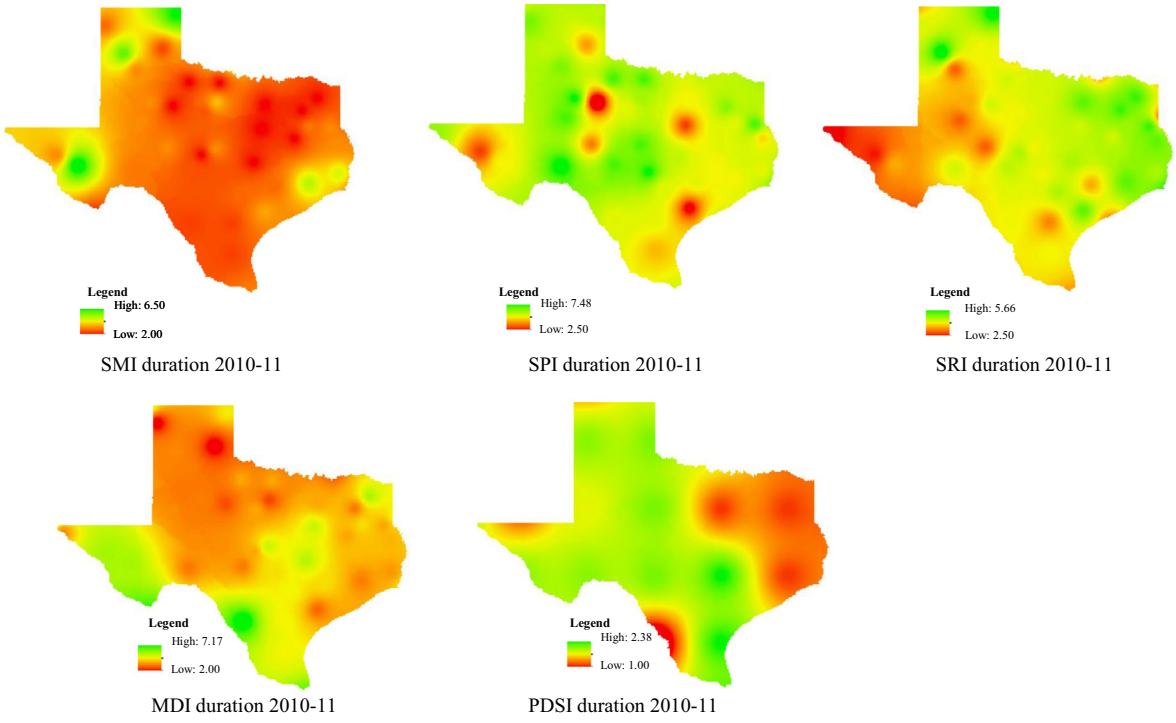

Fig. 7b. Drought duration for Texas during 2010–11 drought using various indices.

followed by drought predictions made using PDSI. The number of drought events predicted by MDI was more (16 for 1950–57 drought and 7 for 2010–11 drought), which shows that the index is capable of capturing finer fluctuations in the drought events. In the case of 2010–11 drought, it can be seen from Table 5 that the maximum drought severity was predicted by PDSI as −14.12, closely followed by MDI which quantified a maximum drought severity of −13.73. The severity levels of 2010–11 droughts were at par with the 1950–57 drought, even though the former event lasted for a much shorter time period. From Figs. 6a and 6b, it can be seen that on an average, MDI predicted higher severity levels in continental, arid, and parts of semi-arid climate zones during

Table 5
Summary of drought properties using different indices during 1950–57 and 2010–11.

Drought properties	SMI	SPI	SRI	MDI	PDSI
1950–1957					
No. of drought events	4	10	9	16	13
Minimum duration (months)	2.20	2.33	2.00	3.00	2.29
Maximum duration (months)	4.37	4.49	5.18	10.45	8.28
Minimum severity	−5.84	−5.02	−10.11	−13.73	−10.01
Maximum severity	−2.52	−2.58	−2.42	−3.29	−4.20
2010–2011					
No. of drought events	3	3	4	7	5
Minimum duration (months)	2.00	2.50	2.50	2.00	1.00
Maximum duration (months)	6.50	7.48	5.66	7.17	2.38
Minimum severity	−5.89	−3.29	−2.38	−2.13	−1.77
Maximum severity	−1.21	−10.73	−10.39	−13.64	−14.12

1950–57 drought, and longer duration levels along arid, semi-arid, and parts of continental and humid regions. In the case of 2010–11 drought, MDI predicted higher severity levels in arid and semi-arid regions, and longer durations in arid and parts of semi-arid region.

5.3. Choice of scale for MDI

MDI is a multiscalar index like SPI. This means that an n-month MDI provides a comparison of drought variables (precipitation, runoff, evapotranspiration and soil moisture) over a specific n-month period with the drought variable totals from the same n-month period for all the years included in the historical record. For example, a 3-month MDI at the end of February compares the December–January–February drought variable total in that particular year with the December to February drought variable totals of all the years. Although the presence of multiple time scales is a distinct advantage, the decision maker may find it difficult to choose a suitable time scale. Only some general guidelines are available to choose between small and large time scales for drought index calculation. Typically, shorter time scales are suitable for analyzing the impacts of agricultural and meteorological droughts and larger time scales are better for hydrological droughts (Mishra and Singh, 2010). Since this is a generalized guideline, the concept of entropy can be applied to help the user choose a suitable scale.

In the methodology section, the basic entropy concepts are discussed. Eq. (5) gives the expression for Shannon entropy that measures the information content that can be obtained from a time series. Higher the value of entropy, higher will be the information content that can be obtained from the time series. By comparing the entropy values of various scales of MDI at a desired location for a desired time period, one can choose the MDI time scale which corresponds to the maximum entropy value. Fig. 8 shows the entropy maps of MDI time scales 1-month, 3-months, 6-months and 12-months for the time period 1950–2012 in Texas. Assuming that the user is interested in the location whose latitude is 30.95° and longitude is −104.75°, and wants to choose a shorter time scale for MDI for further analysis, comparison of entropy maps for 1-month and 3-month MDI can be made. For the given location, the entropy values are 3.56 and 3.74, for 1-month MDI and 3-month MDI respectively. Since 3-month MDI has a higher entropy value for that location, the user may choose it over the 1-month MDI.

6. Conclusion

Due to the presence of multiple drought types which may or may not affect an area simultaneously, the need for a multivariate drought index is essential to jointly represent all the different forms of drought. A new multivariate, multiscalar drought index

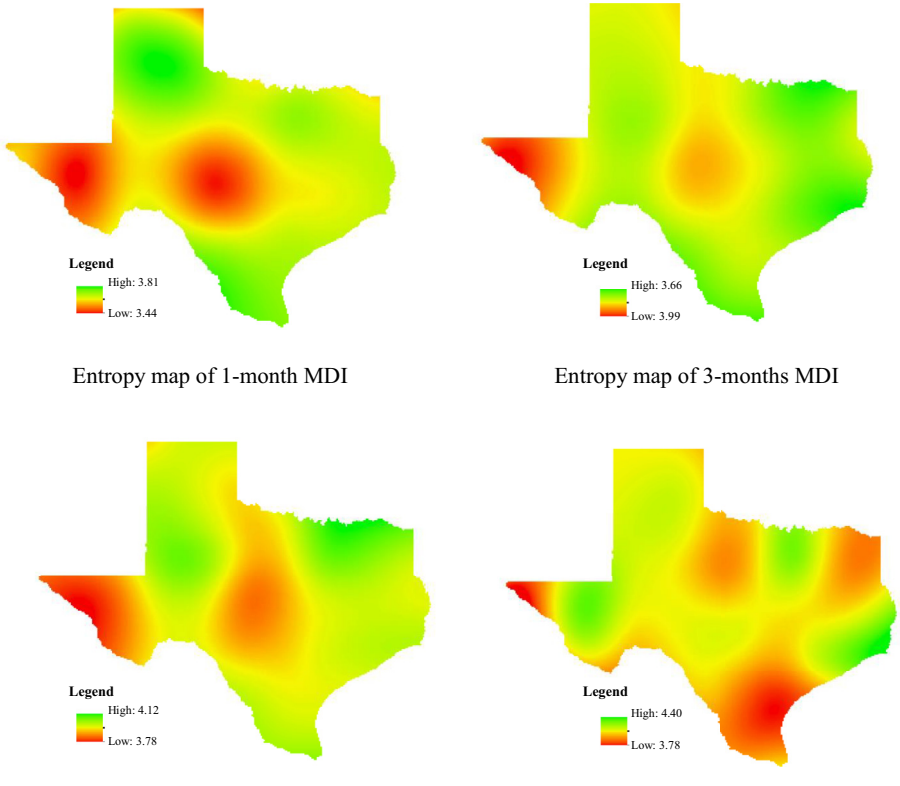

Fig. 8. Entropy maps for various timescales of MDI.

has been developed in this study, based on entropy theory. This study considers all the variables (i.e. precipitation, runoff, soil moisture and evapotranspiration) involved in the water balance for the calculation of a unified drought index named MDI. It is unique in the sense that it accounts for all the physical forms of drought, thus bringing in a broader perspective for drought quantification. MDI was found to be competent in capturing the onset, persistence and termination of droughts. It has several advantages like:

(1) It is multivariate and takes into account the causative variables for different types of drought, thus aggregating the effects of multiple drought forms,
(2) There is a strong information theory background behind the formulation of the index,
(3) The index does not assume linearity between the variables involved,
(4) It can be developed for multiple time scales,
(5) It is statistically robust with a clear and simple computational procedure, and is more flexible and stable for higher dimensional cases when compared to copula based multivariate drought indices, and
(6) It has early drought detection property, and reflects the persistence of drought conditions better than univariate indices.

A simple thumb rule has been suggested based on the entropy value of the drought index time series. For a given location and time frame, the MDI time scale that gives the highest entropy value can be chosen, since a higher entropy value indicates a higher information content.

References

Abramovitz, M., Stegun, I.A., 1964. Handbook of Mathematical Functions, National Bureau of Standards. Applied Mathematics Series-55. Dover, New York, USA.

American Meteorological Society, 2004. Statement on Meteorological Drought. Bull. Am. Meteorol. Soc. 85, 771–773.

Beirlant, J., Dudewicz, E.J., Gyrfi, L., Van der Meulen, E.C., 1997. Nonparametric entropy estimation: an overview. Int. J. Math. Stat. Sci. 6, 17–40.

Benke, A.C., Cushing, C.E., 2005. Rivers of North America. Elsevier Academic Press, Burlington, Massachusetts, pp. 1–1144.

Blumenstock, G., 1942. Drought in the United States Analyzed by Means of the Theory of Probability, Tech. Bul. 819, United States Department of Agriculture, 66, Washington DC, USA.

Borg, I., Groenen, P.J., 2005. Modern Multidimensional Scaling: Theory and Applications. Springer Series in Statistics. Springer.

Bureau of Economic Geology, 1996. River Basin Map of Texas, University of Texas, Austin, Texas, USA.

Dionisio, A., Menezes, R., Mendes, D.A., 2007. Entropy and uncertainty analysis in financial markets. Appl. Phys. Finance Anal., 4–7 July, Portugal.

Ebrahimi, N., Maasoumi, E., Soofi, E.S., 1999. Ordering univariate distributions by entropy and variance. J. Econom. 90 (2), 317–336.

Gringorten, I.I., 1963. A plotting rule for extreme probability paper. J. Geophys. Res. 68 (3), 813814.

Hao, Z., AghaKouchak, A., 2013a. A non-parametric multivariate multi-index drought monitoring framework. Sp. Issue Adv. Drought. Monitoring., Am. Met. Soc. 15, 89–101.

Hao, Z., AghaKouchak, A., 2013b. Multivariate standardized drought index: a parametric multi-index model. Adv. Wat. Res. 57, 12–18.

Heim, R.R., 2002. A review of twentieth-century drought indices used in the United States. Bul. Am. Met. Soc. 83 (8).

Huang, J., van den Dool, H.M., Georgarakos, K.P., 1996. Analysis of model-calculated soil moisture over the United States (1931–1993) and applications to long-range temperature forecasts. J. Clim. 9 (6), 1350–1362.

Jenssen, R., 2010. Kernel entropy component analysis. IEEE Trans. Pattern Anal. Mach. Intell. 32 (5), 847–860.

Jenssen, R., Erdogmus, D., Hild, K.E., Principe, J.C., Eltoft, T., 2005. Optimizing the cauchy-schwarz pdf divergence for information theoretic, non-parametric clustering. In: Proc. Int'l. Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, 34-45, St. Augustine, USA.

Jenssen, R., Erdogmus, D., Hild, K.E., Principe, J.C., Eltoft, T., 2006. Some equivalences between kernel methods and information theoretic methods. J. VLSI Sig. Processing 45, 49–65.

Kao, S.C., Govindaraju, R.S., 2010. A copula-based joint deficit index for droughts. J. Hydrol. 380 (1-2), : 121–134.

Keyantash, J.A., Dracup, J.A., 2004. An aggregate drought index: assessing drought severity based on fluctuations in the hydrologic cycle and surface water storage. Wat. Resour. Res. 40 (9).

Kogan, F.N., 1995. Application of vegetation index and brightness temperature for drought detection. Adv. Space Res. 15 (11), 91–100.

Kramer, M.A., 1991. Nonlinear principal component analysis using autoassociative neural networks. AIChE J. 37 (2), 233–243.

Krause, P., Boyle, D.P., Bse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5 (5), 89–97.

Lathi, B.P., 1968. An introduction to Random Signals and Information Theory. International Textbook Company, Scanton, Pennsylvania.

Liang, X., Lettenmaier, D.P., Wood, E.F., 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., Raschke, E., 1996. A large-scale horizontal routing model to be coupled to land surface parametrization schemes. Tellus 48 (A), 708–721.

Lohmann, D., Raschke, E., Nijssen, B., 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.

McKee, T.B., Doesken, N.J., Kleist, J., 1993. The relationship of drought frequency and duration to time scales. Eight Conf. App. Clim.. Anaheim, California, USA.

McQuigg, J., 1954. A simple index of drought conditions. Weatherwise 7 (3), 64–67.

Mishra, A.K., Singh, V.P., 2011. Drought modeling A review. J. Hydrol. 403, 157–175.

Mishra, A.K., Singh, V.P., 2010. A review of drought concepts. J. Hydrol. 391, 202–216.

Modarres, R., 2007. Streamflow drought time series forecasting. Stoch. Env. Res. Risk Assess. 22, 223–233.

Munger, T.T., 1916. Graphic method of representing and comparing drought intensities. Monthly Weather Rev. 44 (11), 642–643.

Nielson-Gammon, J.W., 1995. Changing Climate of Texas. In: Impact of Global warming on Texas. University of Texas Press, Austin, TX, pp. 39–68.

Palmer, W.C., 1965. Meteorological drought, US Department of Commerce, Weather Bureau, Washington, DC, 58.

Palmer, W.C., 1968. Keeping track of crop moisture conditions, nationwide: the new crop moisture index. Weatherwise 21 (4), 156–161.

Rajsekhar, D., Mishra, A.K., Singh, V.P., 2013. Regionalization of drought characteristics using an entropy approach. J. Hydrol. Eng. 18 (7), 870–887.

Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290 (5500), 2323–2326.

Saul, L.K., Weinberger, K.Q., Ham, J.H., Sha, F., Lee, D.D., 2006. Spectral methods for dimensionality reduction. Semisupervised Learning, 293–308.

Scholkopf, B., 2000. Statistical Learning and Kernel Methods.

Scholkopf, B., Smola, A.J., 2002. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT press.

Scholkopf, B., Smola, A., Mller, K.R., 1999. Kernel principal component analysis. In: Adv. Kernel Methods-Support Vector Learning.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423.

Shepard, D.S., 1984. Computer mapping: the SYMAP interpolation algorithm, Spatial statistics and models, In: Gaile, G.L., Wilmott, C.J., (Eds.), Dordrecht, Netherlands, pp. 133–145.

Shukla, S., Wood, A.W., 2008. Use of a standardized runoff index for characterizing hydrologic drought. Geophys. Res. Lett. 35, L02405.

Svoboda, M., LeComte, D., Hayes, M., Heim, R., Gleason, K., Angel, J., Stephens, S., 2002. The drought monitor. Bul. Am. Met. Soc. 83 (8).

Vicente-Serrano, S.M., Beguera, S., Lpez-Moreno, J.I., 2010. A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. J. Clim. 23 (7).

Wilhite, D.A., 2000. Drought as a natural hazard: concepts and definitions. In: Wilhite, D.A. (Ed.), Drought: A Global Assessment, Hazards Disasters Ser, vol. 1. Routledge, New York, USA, pp. 3–18.

Williams, C.K.I., 2002. On a connection between kernel PCA and metric multidimensional scaling. Mach. Learn. 46, 11–19.

www.hydro.washington.edu.

Yevjevich, V., Siddiqui, M.M., Downer, R.N., 1967. Application of runs to hydrologic droughts. Proceedings of International Hydrology symposium, vol. 1(63). Fort Collins, Colorado, USA, pp. 496–505.