PROBABILISTIC ASSESSMENT OF DROUGHT CHARACTERISTICS USING A HIDDEN MARKOV MODEL

[1]Ganeshchandra Mallya, [2]Shivam Tripathi, [3]Sergey Kirshner, and [4]Rao S. Govindaraju

ABSTRACT

Droughts are characterized by drought indices that measure the departure of meteorological and hydrological variables, such as precipitation and streamflow, from their long-term averages. While many drought indices have been proposed in the literature, most of them use pre-defined thresholds for identifying drought classes ignoring the inherent uncertainties in characterizing droughts. This study employs a hidden Markov model (HMM) for probabilistic classification of drought states. Apart from explicitly accounting for the time dependence in the drought states, the HMM-based drought index (HMM-DI) provides model uncertainty in drought classification. The proposed HMM-DI is used to assess drought characteristics in Indiana using monthly precipitation and streamflow data. The HMM-DI results were compared to those from standard indices and the differences in classification results from the two models were examined. In addition to providing probabilistic classification of drought states, the HMM model is suited for analyzing spatio-temporal characterization of droughts of different severities.

[1] Ph.D. student, School of Civil Engineering, Purdue University, West Lafayette, IN 47907, USA.
[2] Assistant Professor, Department of Civil Engineering, Indian Institute of Technology, Kanpur, UP 208016, India.
[3] Assistant Professor, Department of Statistics, Purdue University, West Lafayette, IN 47907, USA.
[4] Professor, School of Civil Engineering, Purdue University, West Lafayette, IN 47907, USA.

1. Introduction

Droughts are amongst the world's costliest disasters with an annual cost estimated in the range of $6 - $8 billion (Federal Emergency Management Agency 1995). Unlike other natural disasters such as floods and earthquakes, droughts are more difficult to detect because they creep slowly and are already a serious threat before they are detected. Droughts have major impacts on agriculture, natural habitats and ecosystems, and economies of affected regions. Modeled results from climate change scenarios indicate that droughts are likely to intensify over many parts of the world in the next 20-50 years (Dai 2011), suggesting the need for more granularity in drought classification to assess drought impacts accurately for appropriate mitigation strategies.

Based on the data used for the analyses, droughts have been typically classified as meteorological, agricultural, and hydrological droughts (Dracup et al. 1980). Detailed reviews on different types of droughts, their definitions, and the indices used to characterize droughts are available in the literature (Dracup et al. 1980; Heim 2002; Mishra & Singh 2010; Dai 2011). Although drought indices use different forms of water deficits to characterize droughts, the results often do not correspond well among the indices owing to the complex physics that involves precipitation, infiltration, evapotranspiration, groundwater, base flow and direct runoff. Since the definition of a drought is very subjective, no single index is able to address all the causes or impacts of droughts. The desirable characteristics of a drought index (Friedman 1957; Heim 2002) are: 1) flexibility to accommodate appropriate timescale to address the problem at hand; 2) ability to measure longer duration droughts; 3) applicability to the problem being studied; 4) ability to utilize long historical records; and 5) being computable at or near real-time basis.

Researchers have addressed the problem of drought assessment using different approaches such as non-linear models, hybrid models and artificial neural networks (Shin & Salas 2000; Kim & Valdés 2003; Mishra et al. 2007). Copulas have been widely used for modeling the joint dependence structure of drought characteristics, namely intensity, duration and severity (Wong et al. 2009; Madadgar & Moradkhani 2011). Kao and Govindaraju (2010) proposed a joint deficit index (JDI) that uses empirical copulas to provide probability-based description of overall drought status. Hao and Singh (2012) proposed the use of entropy theory for constructing bi-variate joint distribution of drought duration and severity, and provided comparisons with a copula-based analysis. To overcome some of the limitations of parametric frequency analysis, Kim et al. (2006) proposed a non-parametric approach for characterizing the joint behavior of droughts.

One drought index that has gained popularity because of its robustness, computational simplicity, and ability to accommodate different time scales is the standardized precipitation index (SPI; McKee et al. 1993). For a specified time window, accumulated historic precipitation time-series data are used to estimate the probability distribution function and the quantiles of precipitation. The SPI drought index is then computed by applying the inverse standard normal distribution function transformation to the quantiles, making the index values distributed according to a standard normal distribution. SPI values are dimensionless with negative values indicating drought conditions, and the magnitudes of their departures from zero indicating drought severity. The index has been useful to the community, but it has some limitations:

- With the use of fixed thresholds, the frequency of occurrence of droughts is the same for all window sizes and for all stations/regions. While it allows comparisons of drought severities for different locations at a given snapshot in time, it cannot identify *drought-prone* areas (Lloyd-Huges & Saunders 2002).

- For a site where precipitation has small variability, even a small difference in precipitation can lead to differing drought classifications (Lloyd-Huges & Saunders 2002).

- Accumulated precipitation values over a specified time window are assumed to be independent while estimating SPI; this may not be true for larger window sizes (greater than

12 months). Droughts do persist over longer time scales (from few months to several years). Though it is reported in the literature that SPI values are not reliable at time-scales longer than 24 months (Guttman 1999), there are several studies that have investigated multi-year droughts (McKee et al. 1993; Vicente-Serrano 2006). The issue of temporal dependence becomes more poignant for droughts longer than a year. In order to work with non-overlapping segments for SPI, the record length is effectively reduced.

Further, no attempts have been made to account for the inherent uncertainties in classifying a drought state using SPI analysis. Drought-affected regions receive aid based on an assessment of the severity of existing droughts. Drought-readiness schemes are also based on drought classification. Just as in case of floods, decision makers are interested in knowing the uncertainty in drought classification. The allocation of resources and response capabilities of communities will benefit from a probabilistic analysis (Hayes et al. 2004). While many sources of uncertainty exist, the need for an index that provides model uncertainty through a probabilistic classification of drought classes was strongly expressed by decision makers and planners in a recent drought workshop (http://drinet.hubzero.org/tags/ddad2011, http://drinet.hubzero.org/resources/354). Most drought indices were designed for assessment of current conditions *only*, and offer limited or no predictive ability. The US Drought Monitor (USDM; Svoboda et al. 2002) is the most popular source for information on current drought conditions. In the USDM, drought severity levels (D0-D4) are based on percentile rankings of various indicators to depict existing drought conditions (http://drought.unl.edu/dm/monitor.html). While the various indices have allowed for characterizing droughts, they are limited by their inability to offer an estimate of uncertainty in classification of drought states. Further, the methods used for drought classification do not account for the temporal dependence in drought states – a limitation that inhibits forecasting capabilities.

The goal of this paper is to propose and evaluate a new drought index based on hidden Markov model (HMM; e.g., Rabiner 1989). The HMM is a statistical model in which the observations from a system are assumed to be conditioned on unobserved or hidden states that follow a Markov process. To conform to the US Drought Monitor, the number of hidden states in the HMM model was set to 11 where droughts are classified into 5 classes (D0-D4) based on SPI thresholds as in Table 1. In addition, the HMM model can identify a normal state (N) and 5 wet states (W0-W4). Similar to the Joint Deficit Index (Kao & Govindaraju 2010), the HMM-based index (HMM-DI) can be generalized to other hydrologic variables.

Given a time window, both SPI and HMM-DI utilize a time series of cumulative values of the hydrologic variable (precipitation or streamflow in this case). The HMM-DI provides probabilistic classification of drought states reflecting model uncertainty. Because of overlapping time intervals, the time series will be correlated for windows greater than one month. As shown by mutual information analysis in Appendix 1, the series of drought states yielded by SPI are dependent. The HMM-DI offers the advantage of engaging this information explicitly and possesses generative capabilities.

The remainder of the paper is organized as follows: first the data used in the study are described. The mathematical formulation of the HMM-DI is presented. The results obtained in this study are discussed along with comparisons to SPI. The strengths and limitations of the proposed HMM-based drought index are listed, and finally the study conclusions are presented.

2. Data Used

Precipitation and streamflow data from the state of Indiana, located in mid-west United States, were used in this study. Indiana has complex climate patterns with distinct seasons - winters are cold, springs are characterized by thunderstorms and tornadoes, summers are very humid with high temperatures, and autumns are sunny with low humidity. Indiana is located within the US Corn belt;

hence agriculture is one of the major contributors to the state's economy and droughts have significant economic and social impact in the state.

Monthly streamflow data were obtained from the United States Geological Survey (USGS). Streamflow measurements are subjected to human interference, and therefore the data contain both regulated and unregulated flow measurements. For this study, only unregulated streamflow data were used for drought analysis. A total of 36 unregulated USGS gauging stations (see also Kao & Govindaraju 2010) were identified for the study area as shown in Figure 1. Monthly mean discharges for all the 36 stations were collected. The record length for each of these stations is more than 50 years.

Precipitation data were obtained from daily surface data set (TD 3200) of co-operative (COOP) stations from National Climatic Data Center (NCDC). A total of 75 COOP stations were available with data record length greater than 50 years. If the data were missing for an entire month, they were replaced by the historic mean of that specific month (Kao & Govindaraju 2010). Monthly average precipitation data for the nine climatic divisions of Indiana (shown in Fig. 1) were also obtained from NCDC.

3. Methodology

3.1 Mathematical formulation: Hidden Markov Model (HMM)

The HMM (e.g., Rabiner 1989) is a statistical model where observations from a system are assumed to be conditioned on the state of the system. The state is hidden (i.e. not observed) and satisfies the Markov property. The HMM was developed in late 1960s and early 1970s for speech recognition, and it has since been used successfully in many applications including hydrology and climate modeling (Thyer & Kuczera 2003; Robertson et al. 2003; Robertson et al. 2004). The mathematical formulation of the HMM used in this work is described in Tripathi and Govindaraju (2009), and is briefly presented in the following paragraphs.

Let the hydrologic variable of interest at time t be denoted by x_t, $t = 1 \ldots N$ $\{x_t \in R$ and $X = [x_1, \ldots, x_N]^T = x_{1:N}\}$. In HMM, the quantity x_t is assumed to depend on the state variable z_t, $\{Z = [z_1, \ldots, z_N]^T = z_{1:N}\}$ that denotes drought states, is hidden, and follows the first order Markov property. The state variable z_t is a discrete random variable with K values (drought states), $\{1, 2, ..., K\}$. The HMM model can be graphically represented as shown in Figure 2. HMM makes three assumptions about the underlying process being modeled:

1. The drought state z_t evolves according to (first-order) Markov property. Given the drought state at the previous month z_{t-1}, the drought states in the current and future months are independent of past drought states $(z_{t-2}, ..., z_1)$, i.e. $P(z_t \mid z_{t-1}, z_{t-2}, ..., z_1) = P(z_t \mid z_{t-1})$. The probabilities $P(z_t \mid z_{t-1})$ are referred to as transition probabilities.

2. Given the current drought state z_t, the monthly observation x_t for that month is assumed to be conditionally independent of the observations or drought states of other months, $P(x_t \mid x_{1:t-1}, z_{1:t}) = P(x_t \mid z_t)$. The probability distributions $P(x_t \mid z_t)$ are referred to as emission distributions.

3. The transition and emission probabilities depend only on the drought states and observations, and not on the time series index of the observation, $P(z_t = k \mid z_{t-1} = j) = P(z_l = k \mid z_{l-1} = j)$ and $P(x_t \le x \mid z_t = k) = P(x_l \le x \mid z_l = k)$.

Further, assuming the number of states K is known *a priori*, the joint distribution over the SPI categories and monthly hydrologic observations decomposes as a product

$$P(z_{1:N}, x_{1:N}) = P(z_1)\prod_{t=2}^{N} P(z_t \mid z_{t-1})\prod_{t=1}^{N} P(x_t \mid z_t). \quad (1)$$

An HMM can be completely described by:

a) The conditional distribution of the hydrologic variable given the drought state, $P(x_t \mid z_t)$ referred to as emission distribution.

b) The conditional distribution of the present drought state given the previous drought state i.e. $P(z_t \mid z_{t-1})$. Because z_t is a K - valued discrete variable, the conditional distribution is given by a $K \times K$ transition matrix A with elements $A_{jk} = p(z_t = k \mid z_{t-1} = j)$.

c) Marginal distribution of the drought state at the first time step, $p(z_1)$ given by K - dimensional vector with $\pi_k = P(z_1 = k)$.

The posterior probability of being in a state $z_t = k$ at time t is given by

$$P(z_t = k \mid X, \lambda) = \frac{\alpha_t(k)\beta_t(k)}{\sum_{j=1}^{K}\alpha_t(j)\beta_t(j)} \quad (2)$$

where $\alpha_t(k) = P(x_1, x_2, \ldots, x_t, z_t = k \mid \lambda)$ and $\beta_t(k) = P(x_{t+1}, x_{t+2}, \ldots, x_N, z_t = k \mid \lambda)$, and λ represents the set of model parameters, namely the parameters of the emission distributions (θ), the transition matrix (A) and the initial distribution of the states (π).

For a drought index, a definition of drought states that remains unaltered irrespective of the location of a drought is desirable. To achieve this property, the following two steps are taken:

a) The data at any desired time scale (from one month out to several years) are transformed to departures from the mean. This step brings the data from different locations to a common baseline for comparison purposes. The HMM model is applied to the transformed data.

b) The probability density function for the emission distribution is chosen to be a Gaussian distribution of the form

$$p(x_t \mid z_t = k) = N(x_t \mid \mu_k, \sigma_k^2). \tag{3}$$

where μ_k and σ_k^2 are the mean and the variance of a Gaussian distribution, respectively. Because the states are hidden (i.e. not observed), the true nature of emission distributions cannot be determined *a priori*. The choice of Gaussian emission distribution is primarily for mathematical convenience. Many complex processes combine to create droughts, and one may expect that their combined influence, expressed through deviations from the mean, to be Gaussian. Finally, if there is no temporal dependence in the drought states, the HMM automatically collapses to a Gaussian mixture model (GMM) for which theories are well developed (Reynolds & Rose 1995). Since the results of the developed drought index are compared with SPI, the number of states (components in the Gaussian mixture) K was set to 11 (D0-D4, N, W0-W4) as described earlier. Unlike SPI where thresholds are fixed (see Table 1) for drought classification, HMM-DI utilizes a data-driven approach to estimate the parameters of the emission distributions. The μ_k's and σ_k's for all the components of the emission distribution were learnt from the data in a maximum likelihood framework using the Baum-Welch algorithm as described in Rabiner (1989).

3.2 Data Preprocessing and Drought Analysis

The first step in computing a drought index is to collect and pre-process the required data for the study area. The record length of monthly precipitation and streamflow data for all the stations is at least 50 years. Time windows of i months, where i is 1, 3, 6 and 12 months, were chosen to represent typical time scales for precipitation and streamflow deficits. Accumulated monthly precipitation and streamflow time-series were computed corresponding to each time window and for each ending month to account for seasonality as in Kao and Govindaraju (2010). For computing standardized indices, the time-series data were then used to estimate the parameters of the best fit Gamma distribution. The cumulative density function (cdf) of the Gamma distribution was standardized using the standard inverse Gaussian function to compute the SPI drought index. As stated earlier, a negative value of SPI indicates drought conditions and the magnitude of its departure from zero indicates the severity of drought. Standardized streamflow index (SSI; Kao & Govindaraju 2010) was computed along similar lines as the standardized runoff index (SRI; Shukla & Wood 2008), using time series of streamflow to estimate the parameters of the best fit Gamma distribution. The cdf of Gamma distribution was also standardized using standard inverse Gaussian function to obtain SSI values.

According to McKee et al. (1993), a drought event begins when SPI takes a value of -1.0 or less and it ends when SPI becomes positive. In this study, since drought classes are defined to coincide with designations in the US Drought Monitor, a drought event would begin when the SPI took a value of -0.5 or less and it ended when SPI was positive. Thus each drought event had a duration defined by its beginning and end, and an intensity for each month the drought event prevailed. Based on the above definition, the following statistics were noted for SPI: the number of drought events; duration of the drought events; and number and average duration of droughts under each drought category (D0-D4).

For HMM-DI computations, the cumulative monthly time-series data for both precipitation and streamflow were transformed to represent departures from the mean. The HMM model was applied to the transformed data to obtain the probabilistic classification of drought states, i.e. for each time step and a specified window size, the HMM yielded a probability value associated with all the 11 states. In this study, a drought event was defined to begin when the sum of posterior probabilities (Equation 2) of being in D0-D4 states was greater than or equal to 50%, and the drought event ended when it was less than 50%. During a drought event, the drought state with the highest probability was selected as the drought category for the time-step. To enable comparisons with SPI, the following HMM-DI statistics were estimated: the number of drought events; the duration of drought events; and number and average duration of droughts under each drought category (D0-D4).

4. Results and Discussions

4.1 Comparison of HMM-based drought index (HMM-DI) and SPI

The HMM-DI and SPI were computed for all the 75 COOP precipitation stations in Indiana for time windows of 1, 3, 6 and 12 months. The 1-month HMM-DI and SPI for station 120132 at Alpine 2 NE, IN for a 5-year block of 1985-89 are shown in Fig. 3a. The HMM-DI provides probabilistic classification of drought states with the height of each bar indicating the probability of a particular drought, whereas SPI provides a discrete classification. Both models classify the drought states into 5 categories-Abnormally dry, Moderate, Severe, Extreme, and Exceptional (D0-D4). These drought categories are represented in the plot using a legend, and the absence of color indicates a no-drought condition. The precipitation data used in computing the drought states are shown as a line plot. Fig. 3a shows that HMM-DI classifies January 1987 as a D2 category drought with probability >55% and as a D1 category drought with probability ~40%, whereas SPI classifies this month to a D2 category drought. For the next month, the HMM-DI classifies February 1987 into D1 category drought with >90% probability and into D2 category drought with ~5% probability, but SPI classifies it as D4 drought. When precipitation increases in March 1987, SPI classifies it as a normal state, thereby indicating complete recovery from D4 drought of the previous month whereas HMM-DI classifies it as D0 (~45%) drought indicating a more gradual recovery. By design, the HMM-DI model accounts for temporal dependence in the drought states explicitly.

The importance of the time dependence built in HMM-DI is even more relevant for larger window sizes (see Appendix 1). Fig. 3b shows a comparison of 3-month HMM-DI and SPI values. The cumulative precipitation for the 3-month window is shown as a solid line. HMM-DI classifies January 1987 as a D1 drought (~55%) and D0 drought (~45%), whereas SPI classifies the same month as D1 category drought. When the precipitation deficit increases in February 1987, HMM-DI classifies it as a D2 category drought (~ 80%) and D3 drought (~20%) whereas SPI shows a sudden transition from D1 to D4 drought. In the following months when the precipitation deficit decreases, HMM-DI is able to capture the gradual transition of drought states. For example, a small increase of precipitation in July 1987 causes SPI to change drought classification from D1 to D0 category drought. HMM-DI, on the other hand, classifies July 1987 as a D1 category drought (~90%) but also shows small signs of recovery to D0 category drought (~10%). With increasing window sizes of 6 and 12 months in Figs. 3c and 3d, respectively, the temporal dependence in the accumulated time series and the drought states also increases. The smoother transitions are reflected in the HMM-DI results.

The HMM-DI was next applied to streamflow data. The transformed streamflow data at the 36 USGS unregulated stations in the study area (Fig. 1) were used to compute the HMM-DI and the results were compared with the standardized streamflow index (SSI). As an example, Fig. 4a compares 1-month HMM-DI with SSI values at USGS station 3275000 at Whitewater River Alpine, IN for a 5-year block of 1985-89. Streamflow values at the station during this time period are plotted as a line plot in the same figure. The stream gauge is located at a distance less than 10 kilometer from the COOP station (120132) used in the foregoing analysis. Streamflow at a gauging site is influenced by many factors including rainfall over the entire contributing area. As found in previous studies (Kao & Govindaraju 2010), the cross-correlation between precipitation and streamflow at these stations was significant to suggest close mapping of temporal dependencies in drought states. Hence, a comparison of drought states observed at COOP station (120132) and USGS station (3275000) is provided here. This USGS streamflow gauging station is characterized by high spring flows and low fall flows. The streamflow during January 1988 was very low compared to the long-term January mean flow. In Fig. 4a, HMM-DI estimates a D3 category drought (probability ~80%) during January 1988 and with a smaller probability (~20%) of a D2 category drought. The SSI estimates the drought to be of D1 category. Comparison of HMM-DI at the USGS station and at nearby COOP station shows that during August- November 1987 the COOP station experienced a dry spell, resulting in meteorological drought of D1-D3 category (Fig. 3a). This rainfall deficit likely had an impact on the streamflow, thus resulting in extended and severe hydrologic drought between November 1987 to July 1988.

For window sizes greater than 1-month, cumulative streamflow data were used and then transformed to represent deviations from the mean. This transformed data was then used to compute SSI and HMM-DI values. For a 3 month time window, the HMM-DI classifies February-April months in 1988 as D1 category drought (Fig. 4b) that can be attributed to longer memory in the data coupled with the time dependence built in the model. Figs. 4c and 4d show the HMM-DI and SSI comparisons at the selected USGS station for time windows 6 and 12 months respectively. As expected, the streamflow data are smoother, suggesting gradual transitions in drought states that are better reflected in the higher granularity provided by the HMM-DI.

4.2 Comparing HMM-DI and SPI Statistics

From the analysis of past drought records, it is expected that the number of extreme (D4, D3) drought events would be smaller for larger window sizes, and that the duration of a drought increases with window size. Since SPI classification is based on predefined thresholds, the number of drought events and their durations is of the same order irrespective of window size. This is evident in Fig. 5a, where boxplots of average duration of D2 category droughts are compared for various window sizes for all 75 precipitation COOP stations. While the average duration of D2 category drought is of the same order irrespective of window sizes for SPI, the average duration increases with increase in window size for HMM-DI. Similarly, Fig. 5b shows the boxplot of the number of D2 category drought events versus window size. For SPI, the numbers of D2 category drought events for window sizes 3 – 12 months are approximately the same. HMM-DI on the other hand shows a stronger trend of number of events decreasing with increase in window size. Figure 5c compares relative frequency of D2 category droughts for different window sizes for all the 75 precipitation COOP stations using boxplots. Because of the specification of thresholds used in SPI as in Table 1, the relative frequency for a given drought category is preordained to be the same for different window sizes. In Figure 5c, the relative frequency of occurrences of D2 droughts for different window sizes for all the 75 COOP stations is constant (~4%) according to SPI. However, the HMM-DI does show an increase in the relative frequency of D2 droughts with increasing window size. The HMM-DI statistics are likely to be sensitive to the strategy adopted for counting the numbers and durations of droughts. However, the qualitative behavior of HMM-DI in revealing the trends as shown in Fig. 5 is consistent. Similar trends were observed for other drought categories, but only the results obtained for D2 category are described here for brevity.

4.3 Comparison of emission distributions over climate divisions at different time scales

Emission distributions reflect the nature of droughts over a region as revealed by the data. As an example, emission distributions of drought states (D0-D4) for 1-month time window are shown in Fig. 6a for rainfall data that were aggregated over climatic divisions 1 and 9 (see Fig. 6). These two divisions have the greatest geographical separation. The probability density functions (pdf) for the entire data was determined by a non-parametric kernel density estimation method (Bowman & Azzalini 1997) and is shown as a thick black line for the two climatic divisions in Fig. 6a. While both these pdfs are positively skewed, the pdf for climatic division 1 has a steeper rising limb with more probability mass in the range corresponding to droughts thereby indicating a higher propensity for droughts in this division.

Further, apart from D2 category, the emission distributions for the droughts classes are more peaked and less diffuse (smaller variance) in climatic division 9. Thus, droughts in this division are classified with higher probabilities, and there would be less uncertainty in the determination of drought category for 1-month precipitation data. The proposed HMM-DI utilizes information from the emission distributions contained in the precipitation data. However, this information is not engaged in drought classification by HMM-DI with fixed emission distributions (Mallya et al. 2010).

The emission distributions of the drought states (D0-D4) for 3, 6 and 12 months time windows (Fig. 6b - 6c) are compared between climate divisions 1 and 9. As in Fig. 6a for a 1-month drought, Fig. 6b shows that pdfs are consistently negatively skewed with a steeper rising limb for climatic division 1 than for division 9. In contrast to Fig. 6a, the emission distributions for larger time windows and for the various drought categories are more peaked with smaller variances for climatic division 1. Moreover, for 3 and 12-month drought windows (Fig. 6b and 6d), the emission distributions for moderate drought categories D1 and D2 are similar for these two regions despite their geographical separation – suggesting that probability of droughts in these categories (D1 and D2) tend to be similar in both divisions 1 and 9. Similarly, for 6-month drought windows (Fig. 6c), the emission distributions for D1, D3 and D4 categories are similar when compared between divisions 1 and 9. The emission distributions are thus useful for analyzing the nature of droughts over a region. As the time scale of a drought increases, the emission distributions for all drought categories tend to have smaller peaks and the variances increase.

4.4 Emission distributions for streamflow

Emission distributions of drought states (D0-D4) for 1, 3, 6 and 12 months are shown in Fig. 7 for streamflow data at USGS station 3275000 at Whitewater River Alpine, IN. Similar to the foregoing analysis, the probability density function (pdf) for the streamflow series was determined by a non-parametric kernel density estimation method, shown as a black thick line. While pdfs for 1, 3 and 6 months time window (Fig. 7a-7c) are positively skewed with steep rising limb, the pdf for 12-months time window shows a bi-modal distribution at this station. Further, the probability mass for these pdfs is higher in the region corresponding to droughts, again indicating a higher drought propensity following the trend in precipitation. For 1-month time window (Fig. 7a), the emission distribution for D0 category drought has the highest peak and also a small variance. This indicates that hydrologic droughts at this station are likely to be classified in D0 category with higher probabilities and with less uncertainty. For larger time windows of 3, 6, and 12 months, the emission distributions for severe category droughts (D2 and D3; Fig. 7b-7d) have higher peaks suggesting that severe droughts are more likely at longer time scales at this station. Because streamflows represent aggregated response from a contributing watershed, a direct comparison over different stations is not meaningful. With Fig. 7, we demonstrate the role of emission distributions at a location, but these distributions are reflective of hydrologic processes over the corresponding watershed. From Figs. 6 and 7, emission distributions are shown to be useful tools in reflecting the nature of droughts at a location. Given that data are accumulated over a given time window for HMM-DI computations, it is not proper to compare emission distributions for different time windows as they are not normalized with respect to each other. This aspect is also reflected in the horizontal axis scale in Figs. 6a-d where the magnitudes of the deviations increase with increasing time windows.

5. Model and data limitations

The proposed HMM-DI model $(\lambda = \theta, A, \pi)$ requires parameters of the emission distributions, elements of the transition matrix and the initial distribution of states. Therefore, if one were to choose the Drought Monitor classification scheme that involves 11 states (D4-D0, N, and W0-W4), the total number of free parameters that need to be estimated would be equal to 142 (i.e. 11×2 + 11×10 + 10). There is no known analytical solution to this problem even if a finite observation sequence were given as a training data. The standard approach is to estimate the parameters using the Baum-Welch method (Rabiner 1989) such that the probability of observation, given the model, is maximized. The maximum likelihood estimate is a 'point' estimate. Similar to other iterative methods, the Baum-Welch algorithm yields a local maximum in the likelihood surface. Consequently, a hundred random initial estimates were tried in search of stable results. The log-maximum likelihood functions for each initialization were compared to test if indeed the global maxima was reached, and if the corresponding final parameter estimates were consistent. It was observed that the final parameter estimates for the mean and standard deviation of the emission distributions corresponding to each of the eleven components of the HMM-DI were not always consistent, likely due to high dimensionality of maximum likelihood surface resulting from the large set of parameters that need to be estimated from a finite observation sequence. If the number of components of the HMM model are reduced, then the dimensionality of the parameter space also reduces substantially, leading to more consistent estimates of parameters. For instance, when the numbers of components are reduced to three, the dimensionality of the parameter space is reduced from 142 to 14. After numerous trials (see Mallya 2011), it was found that stable results could be obtained for five components. Thus the accepted standard of 50 years of record length for hydrologic data may not be sufficient for stable results if all extreme drought states of US Drought Monitor are to be adopted for analysis.

If a Gaussian mixture model (GMM) with 11 components is used instead of a HMM based model, the dimensionality of the parameter space is reduced from 142 to 22 because of the absence of the transition matrix for a GMM. Stable results were obtained for emission distribution parameters for this model (Mallya 2011). However, as noted during the exploratory analysis (see Appendix 1), the GMM is useful only for probabilistic drought classification at 1-month time scale. At time scales greater than one month, there is significant dependence between drought states of neighboring months as suggested by the mutual information statistics, and the GMM does not incorporate the dependence in the data.

To obtain a solution for unique parameter estimates for the HMM model, while preserving the dependence in the drought states and also maintaining eleven components in order to be consistent with the Drought Monitor classification scheme, a modified transition matrix such that only bidiagonal elements are updated during each iteration, was explored. The probability of drought states changing by more than one category in either direction in a one month time frame is small. This is especially true for longer duration droughts which have strong dependence built-in due to overlapping data. This smoother approach is suggested as an alternative when the data record length is small and data sufficiency is an issue. By adopting this approach, the dimensionality of the parameter space may be reduced from 143 to 55, resulting in an improved maximum likelihood surface. The experiment was performed for longer duration droughts (greater than 1-month). Stable parameter estimates were obtained for drought durations 9-months or greater, with eleven components in the HMM model. If the number of drought classes were reduced, stable results were obtained even for smaller duration droughts as shown in Mallya (2011).

A note of caution is needed when missing data are replaced with their long-term mean - a standard practice in hydrology. If data are missing for several months, then artificially more mass is placed on one of the components of the emission distribution with zero mean and negligible standard deviation. This problem is particularly relevant when analyzing 1-month duration droughts. At longer time windows, the problem is muted because of overlap from neighboring months.

6. Summary and conclusions

A hidden Markov model was used to develop a new drought index. The parameters of the HMM were estimated using the method of maximum-likelihood. The developed drought index (HMM-DI) was applied to precipitation and streamflow data over Indiana and compared with the standardized indices (SPI and SSI). The HMM-DI explicitly incorporates temporal dependence in drought states, and consequently in cumulative rainfall amounts over all time windows. The emission distributions provide an opportunity for examining the distributional properties of droughts of different severities. The important conclusions from this study are as follows:

1. The HMM-DI provides a smooth transitioning of drought severity over time. This probabilistic classification provides a more informed measure of drought severity allowing for better mitigation measures.

2. The average duration of a drought in any category increases with the size of the time window and is revealed clearly by HMM-DI.

3. The HMM-DI shows that emission distributions for severe droughts tend to be similar across climatic divisions for longer duration droughts.

Since drought indices are designed with a specific purpose depending on local and regional needs, there is no one true index. Rather than pre-defined thresholds, the HMM-DI allows the data to determine classification boundaries, and provides new insights into drought features. Comparisons with classifications from SPI or SSI were only for revealing differences in results between the models.

The current study evaluated data at individual locations or aggregated data over climatic divisions. The graphical nature of this index could be exploited to provide a principled approach for searching physical mechanisms that trigger droughts. Given the generative nature of HMM-DI, it can be used for short-term drought forecasting, computation of future water deficits, and for estimating the probability of recovering from existing droughts.

The HMM model can be developed in an online (parameters estimated adaptively) or offline (static parameters) mode. This paper presented results in an offline mode to enable comparisons with standardized indices. An online model would be useful for operational purposes and would go hand in hand with examination of generative properties of the model. These topics will form the basis for future studies.

7. Acknowledgments

Studies of the authors were supported in part by the National Science Foundation under Grants DBI 0619086, OCI 0753116, and AGS 1025430. This support is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix 1

A.1 Exploratory Analysis

Exploratory analysis was carried out to check if the drought states adopted for computing SPI were independent in time as implicitly assumed in its formulation. Because there is an overlap in the data series adopted by SPI at drought time-windows greater than 1-month, it is important to check the extent of this temporal dependence. In this regard, the impetus for developing an HMM-type model is the use of latent variables (i.e. drought states) that allow for a natural way of incorporating uncertainty in classification of drought states.

Mutual information is a measure of mutual dependence between random variables. This dimensionless quantity expresses the uncertainty measure of one random variable given the knowledge of the rest. High mutual information indicates reduced uncertainty; low mutual information indicates high uncertainty; and zero mutual information indicates that the two random variables are independent. Given two discrete random variables *X* and *Y* with a joint probability distribution $P_{X,Y}(x, y)$, mutual information $I(X;Y)$ is defined as follows (Shannon & Weaver 1949; Cover & Thomas 2001):

$$I(X;Y) = \sum_{x,y} P_{XY}(x, y) \log \frac{P_{XY}(x, y)}{P_X(x) P_Y(y)} \qquad \text{(A.1)}$$

where $P_X(x) = \sum_y P_{XY}(x, y)$ and $P_Y(y) = \sum_x P_{XY}(x, y)$ are the marginal distributions of X and Y. Unlike linear correlation which assumes the joint distribution to be a bivariate Gaussian, equation (A.1) may be applied to any two discrete random variables. The mutual information statistic was computed using equation (A.1) for several combinations of drought categories (or bins) using SPI values at different time scales. For brevity, only results from precipitation station SI124181 are discussed here. Figure 8a shows the mutual information statistic for January month drought states for two bins. The two bins were selected by combining D4-D0 and N-W4 classes, respectively. A mutual information statistic value close to 1 indicates strong dependence in drought classification of one month (e.g., February) given the drought classification of another (e.g., January). Figure 8a shows that there is very little mutual information between two consecutive months at 1-month time scale. This suggests a lack of dependence in the drought states between any two consecutive months, and hence little advantage can be achieved by using HMM model for 1-month drought classification. A simpler Gaussian mixture model (GMM) may be used instead of HMM at this time scale for probabilistic drought classification. By using GMM, the number of model parameters that require estimation can be reduced significantly, thereby improving robustness of parameter estimates and reliability of model results. As expected, the mutual information between different months increases with the time window. For instance, Fig. 8a shows persistent dependence for 12 month time window, and drought states from February to May have high mutual information with the drought state in the January month.

Figure 8b shows the mutual information statistics for January month drought state when D4-D2, D1-D0 and N-W4 categories are combined to form three bins. We observe that with increase in number of bins, the details of mutual information statistic have improved, but again there is lack of mutual information between two consecutive months indicating independence for 1-month droughts, but the mutual information is strong for longer drought windows. This is substantiated further in Figure 8c that shows the mutual information statistics for January drought states with four bins: D4-D2, D1-D0, N, and W0-W4. The results again suggest that there is lack of information between two consecutive months for 1-month droughts, but with increasing time window, there is a strong dependence in drought states. A detailed analysis of the mutual information results for different time windows and for different groupings is provided in Mallya (2011). Thus, a model that preserves temporal dependence is needed for proper characterization of droughts of durations greater than one month. This was achieved using hidden Markov models in this study.

8. References

Bowman, A.W. & Azzalini, A., 1997. *Applied smoothing techniques for data analysis*, Oxford University Press, USA.

Cover, T.M. & Thomas, J.A., 2001. Elements of Information Theory. In *Elements of Information Theory*. John Wiley & Sons, Inc.

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

Dracup, J.A., Lee, K.S. & Jr, E.G.P., 1980. On the definition of droughts. *Water Resources Research*, 16(2), pp.297–302.

Federal Emergency Management Agency, 1995. National Mitigation Strategy. *FEMA Washington, DC.*

Friedman, D.G., 1957. The prediction of long-continuing drought in south and southwest Texas. *Ocassional papers in Meteorology*, 1, p.182.

Guttman, N.B., 1999. Accepting the standardized precipitation index: A calculation algorithm. *JAWRA Journal of the American Water Resources Association*, 35(2), pp.311–322.

Hao, Z. & Singh, V.P., 2012. Entropy based method for bivariate drought analysis. *Journal of Hydrologic Engineering*, 1(1), pp.425–425.

Hayes, M.J., Wilhelmi, O.V. & Knutson, C.L., 2004. Reducing drought risk: bridging theory and practice. *Natural Hazards Review*, 5(2), pp.106–113.

Heim, R.R., 2002. A review of twentieth-century drought indices used in the United States. *Bulletin of the American Meteorological Society*, 83(8), p.1149.

Kao, S.-C. & Govindaraju, R.S., 2010. A copula-based joint deficit index for droughts. *Journal of Hydrology*, 380(1-2), pp.121–134.

Kim, T.-W. & Valdés, J.B., 2003. Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks. *Journal of Hydrologic Engineering*, 8(6), pp.319–328.

Kim, T.-W., Valdés, J.B. & Yoo, C., 2006. Nonparametric approach for bivariate drought characterization using Palmer drought index. *Journal of Hydrologic Engineering*, 11(2), pp.134–143.

Lloyd-Huges, B. & Saunders, M.A., 2002. A drought climatology for Europe. *International Journal of Climatology*, 22, pp.1571–1592.

Madadgar, S. & Moradkhani, H., 2011. Drought analysis under climate change using copula. *Journal of Hydrologic Engineering*, 1(1), pp.357–357.

Mallya, G., 2011. *Hidden Markov model based probabilistic assessment of droughts*. West Lafayette, IN 47907: Purdue University.

Mallya, G., Tripathi, S. & Govindaraju, R.S., 2010. Assessment of drought characteristics using graphical models. In Frontiers of Interface between Statistics and Sciences. Hyderabad, India, pp. 565–575.

McKee, T.B., Doesken, N.J. & Kleist, J., 1993. The relationship of drought frequency and duration to time scales. *Conference of Applied Climatology, American Meteorological Society, Anaheim, CA*.

Mishra, A.K., Desai, V.R. & Singh, V.P., 2007. Drought forecasting using a hybrid stochastic and neural network model. *Journal of Hydrologic Engineering*, 12(6), pp.626–638.

Mishra, A.K. & Singh, V.P., 2010. A review of drought concepts. *Journal of Hydrology*, 391(1–2), pp.202–216.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), pp.257–286.

Reynolds, D.A. & Rose, R.C., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1), pp.72–83.

Robertson, A.W., Kirshner, S. & Smyth, P., 2003. *Hidden Markov models for modeling daily rainfall occurrence over Brazil*, University of California.

Robertson, A.W., Kirshner, S. & Smyth, P., 2004. Downscaling of daily rainfall occurrence over northeast Brazil using a hidden Markov model. *Journal of Climate*, 17(22), pp.4407–4424.

Shannon, C.E. & Weaver, W., 1949. *The mathematical theory of communication.*, Urbana, Illinois, USA: University of Illinois Press.

Shin, H.-S. & Salas, J.D., 2000. Regional drought analysis based on neural networks. *Journal of Hydrologic Engineering*, 5(2), pp.145–155.

Shukla, S. & Wood, A.W., 2008. Use of a standardized runoff index for characterizing hydrologic drought. *Geophysical Research Letters*, 35, p.7.

Svoboda, M. et al., 2002. The Drought Monitor. *Bulletin of the American Meteorological Society*, 83, pp.1181–1190.

Thyer, M. & Kuczera, G., 2003. A hidden Markov model for modelling long-term persistence in multi-site rainfall time series 1. Model calibration using a Bayesian approach. *Journal of Hydrology*, 275(1-2), pp.12–26.

Tripathi, S. & Govindaraju, R.S., 2009. On the identification of intra-seasonal changes in the Indian summer monsoon. In *Proceedings of the Third International Workshop on Knowledge Discovery from Sensor Data*. SensorKDD ’09. New York, NY, USA: ACM, pp. 62–70.

Vicente-Serrano, S., 2006. Differences in spatial patterns of drought on different time Scales: An analysis of the Iberian peninsula. *Water Resources Management*, 20(1), pp.37–60.

Wong, G. et al., 2009. Drought analysis using trivariate copulas conditional on climatic states. *Journal of Hydrologic Engineering*, 15(2), pp.129–141.

List of Tables

List of Figures

	and 12- month(s) time window. The thick black line represents the probability density function (pdf) of the cumulative precipitation in that window.
7	Plot of emission distributions for all drought states (D0-D4) estimated at Whitewater River at Alpine,IN (3275000) USGS streamflow station. The emission distributions correspond to 1, 3, 6 and 12- month(s) time window. The thick black line represents the probability density function (pdf) of the cumulative streamflow in that window.
8a	Mutual information statistics for January month SPI at different time scales (1, 3, 6 and 12 months) using two bins – D4-D0 and N-W4.
8b	Mutual information statistics for January month SPI at different time scales (1, 3, 6, and 12 months) using three bins – D4-D2, D1-D0, and N-W4.
8c	Mutual information statistics for January month SPI at different time scales (1, 3, 6, and 12 months) using four bins – D4-D2, D1-D0, N, and W0-W4.

Tables:

Table 1: US Drought Monitor classification scheme

Category	Description	SPI Range
D0	Abnormally Dry	-0.5 to -0.7
D1	Moderate Drought	-0.8 to -1.2
D2	Severe Drought	-1.3 to -1.5
D3	Extreme Drought	-1.6 to -1.9
D4	Exceptional Drought	-2.0 or less

Figures:

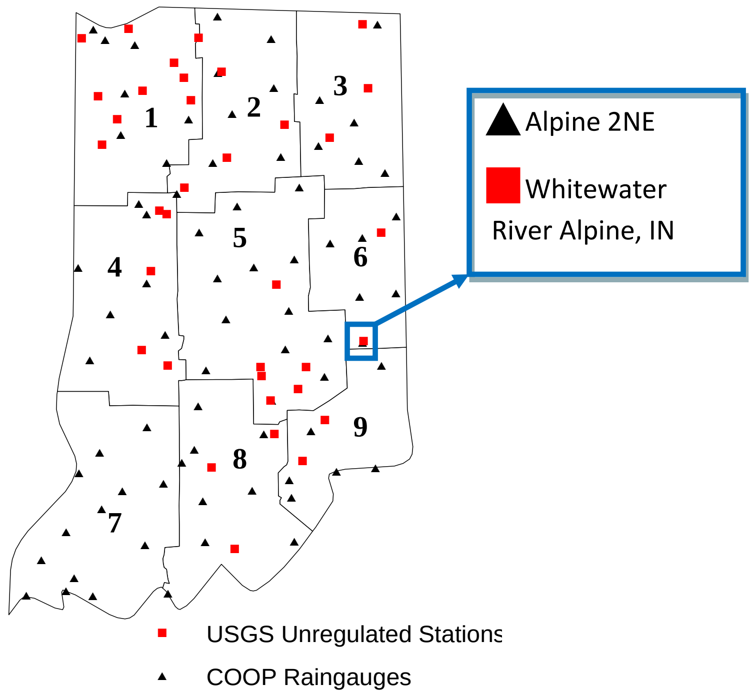


Figure 1: Map showing the study area with the location of COOP raingauges and USGS unregulated stations.

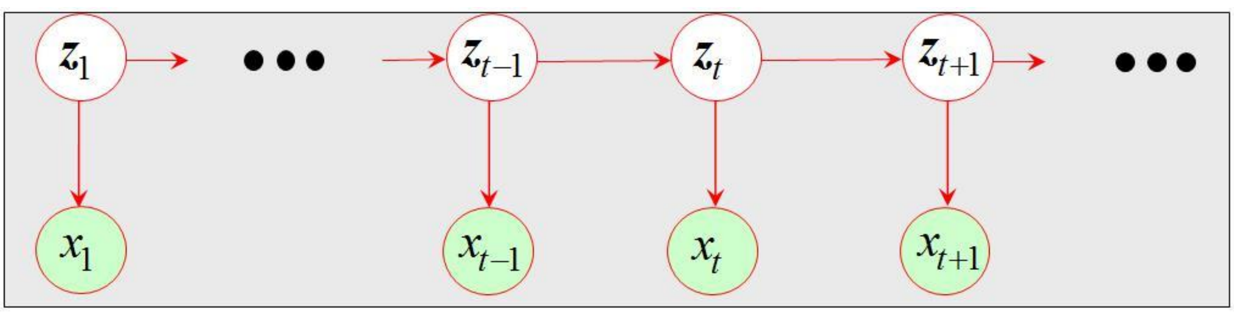
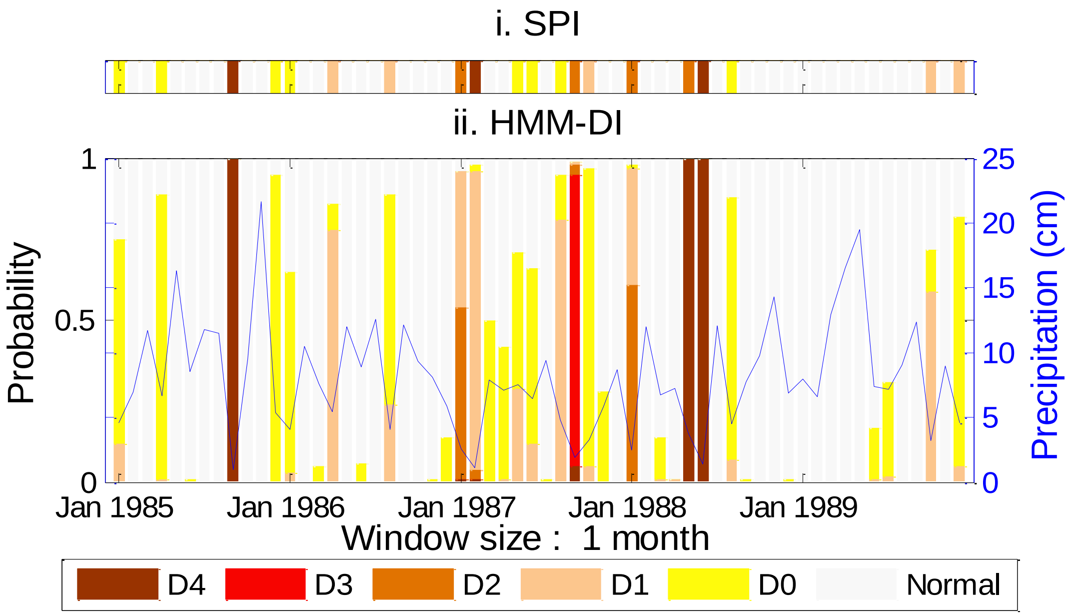
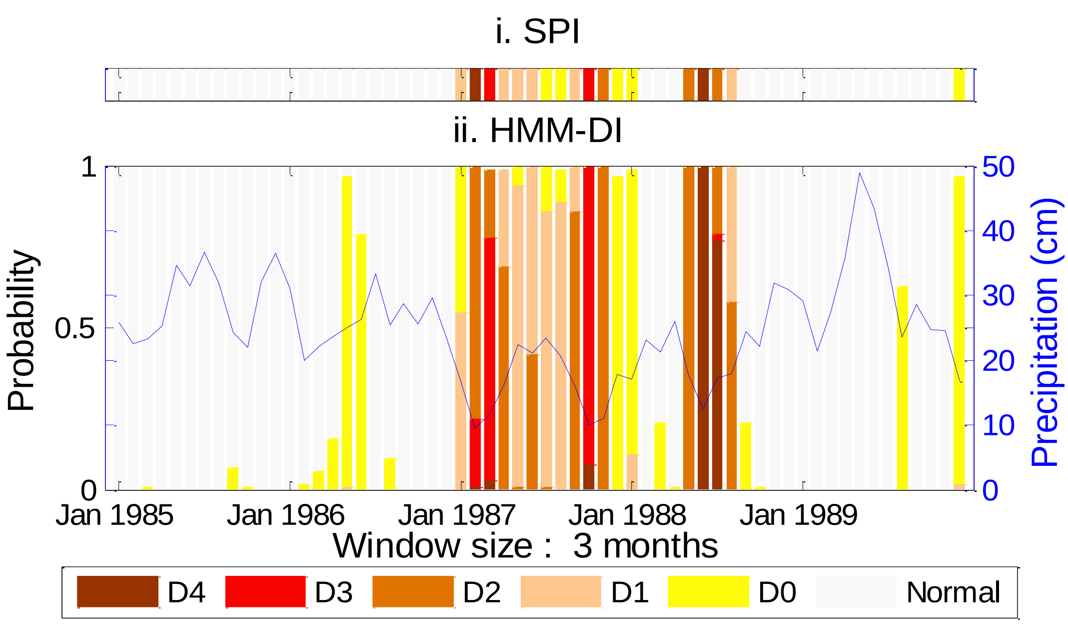


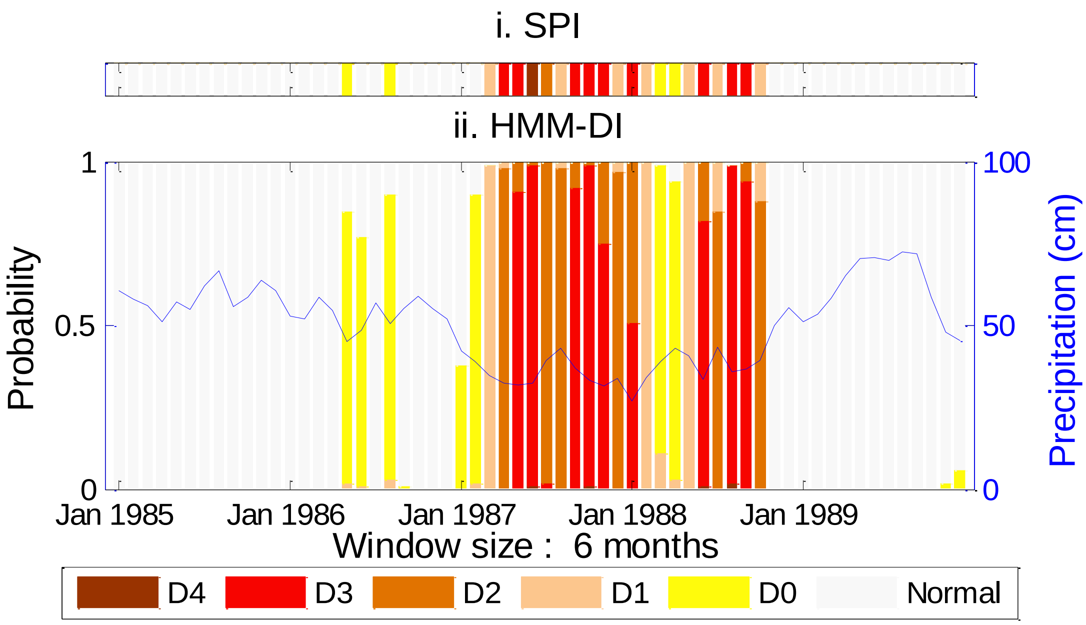
Figure 2: Graphical representation of HMM Model. Here x_t refers to the observed hydrologic time series, while z_t refers to the hidden drought state.


i. SPI
ii. HMM-DI
Probability
1
0.5
0
Precipitation (cm)
25
20
15
10
5
0
Jan 1985
Jan 1986
Jan 1987
Jan 1988
Jan 1989
Window size : 1 month
D4
D3
D2
D1
D0
Normal

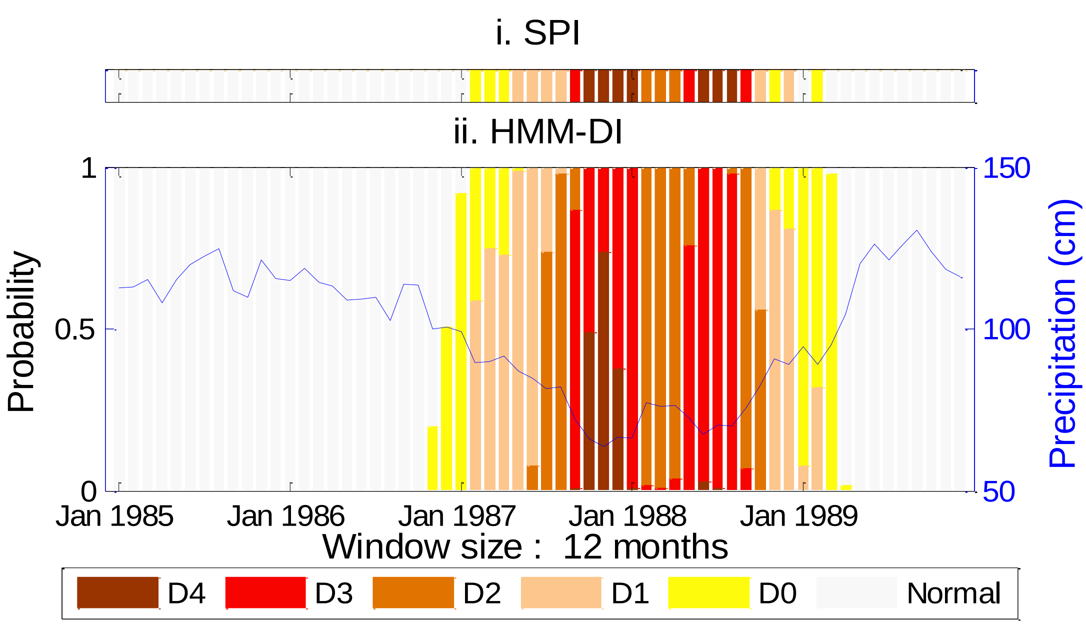
(a)


i. SPI
ii. HMM-DI
Probability
1
0.5
0
Precipitation (cm)
50
40
30
20
10
0
Jan 1985
Jan 1986
Jan 1987
Jan 1988
Jan 1989
Window size : 3 months
D4
D3
D2
D1
D0
Normal

(b)

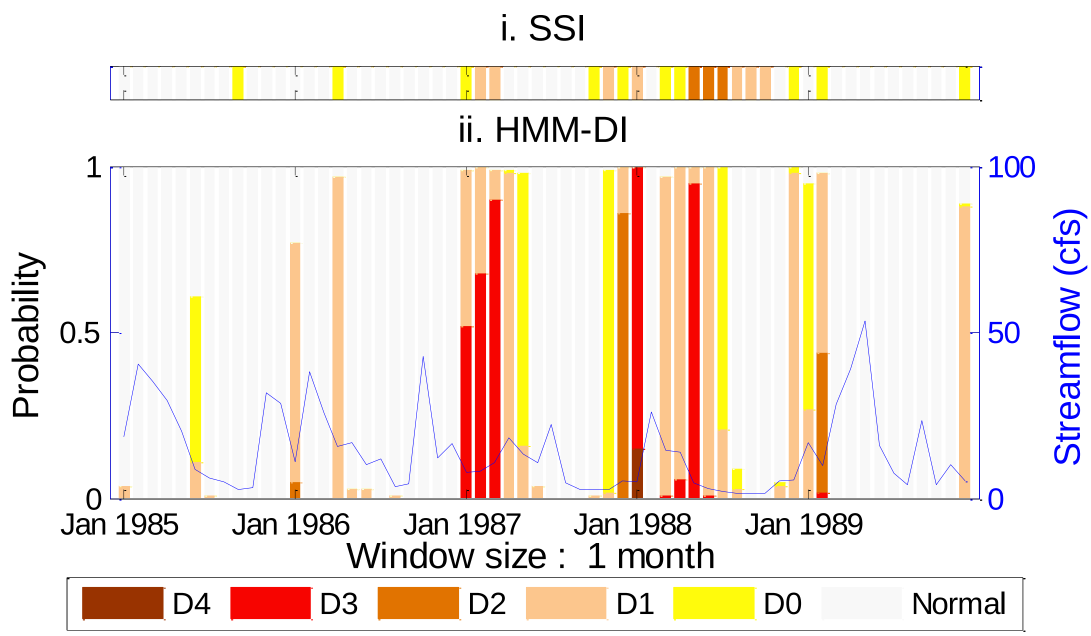


(c)

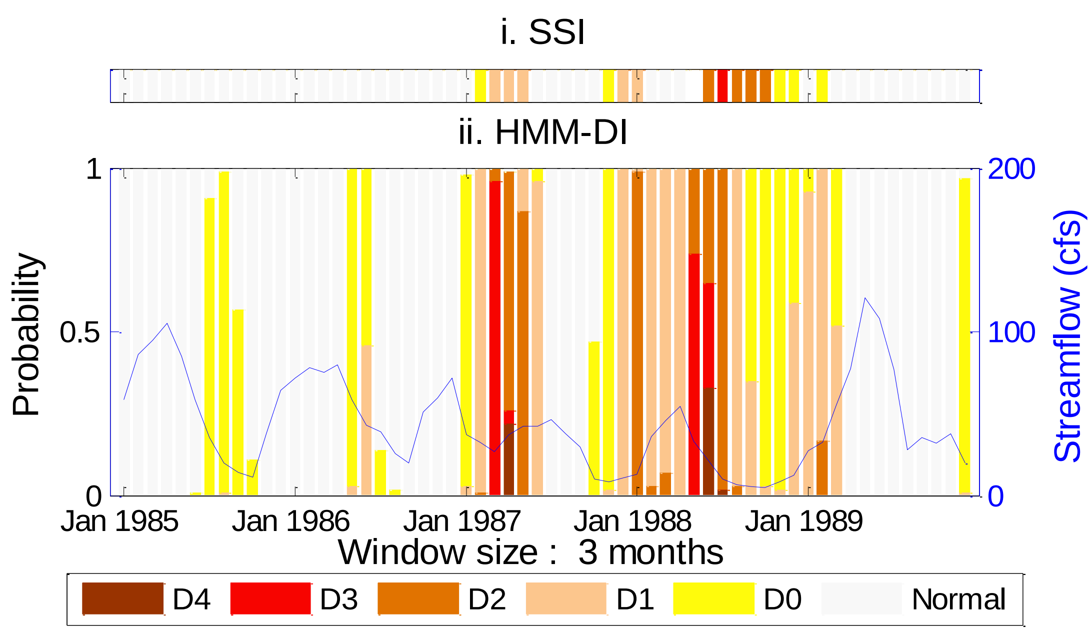


(d)

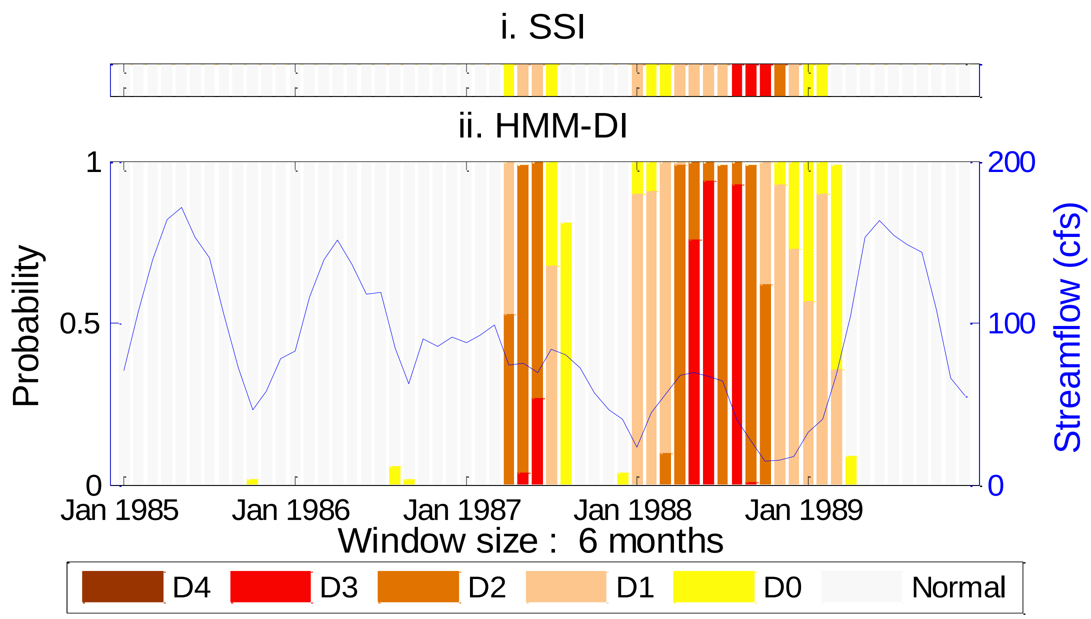
Figure 3: Graph showing HMM-DI and SPI drought classification for a 1,3,6 and 12-month(s) window at Alpine 2 NE station for a 5-year block 1985-89. The line plot corresponds to the cumulative precipitation total at the location used for computing the results based on the window size.


i. SSI
ii. HMM-DI
Probability
Streamflow (cfs)
1
0.5
0
100
50
0
Jan 1985
Jan 1986
Jan 1987
Jan 1988
Jan 1989
Window size : 1 month
D4
D3
D2
D1
D0
Normal

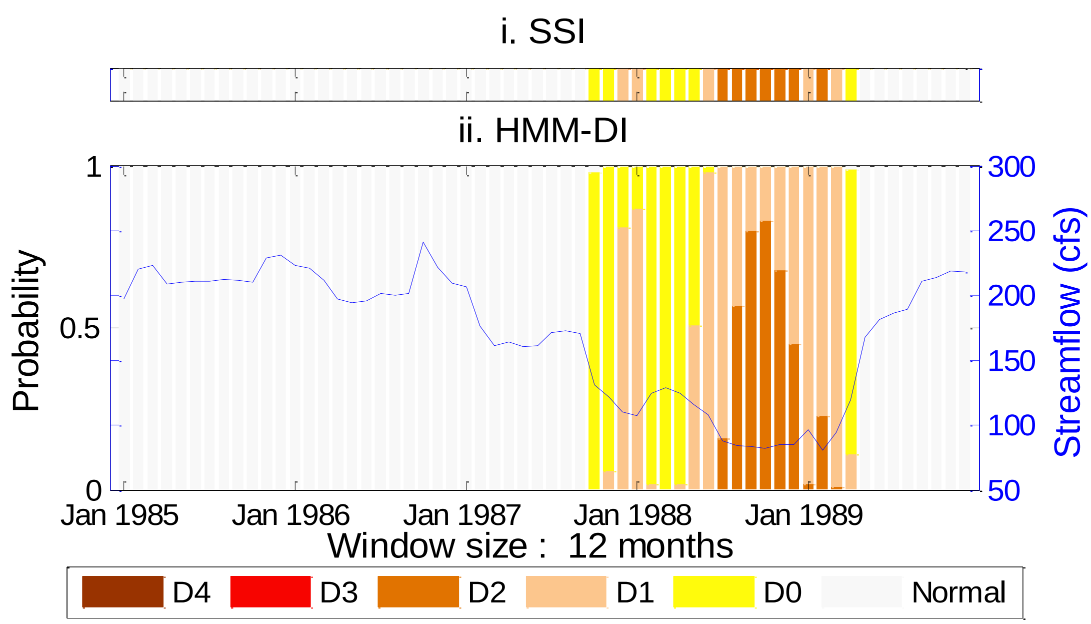
(a)


i. SSI
ii. HMM-DI
Probability
Streamflow (cfs)
1
0.5
0
200
100
0
Jan 1985
Jan 1986
Jan 1987
Jan 1988
Jan 1989
Window size : 3 months
D4
D3
D2
D1
D0
Normal

(b)

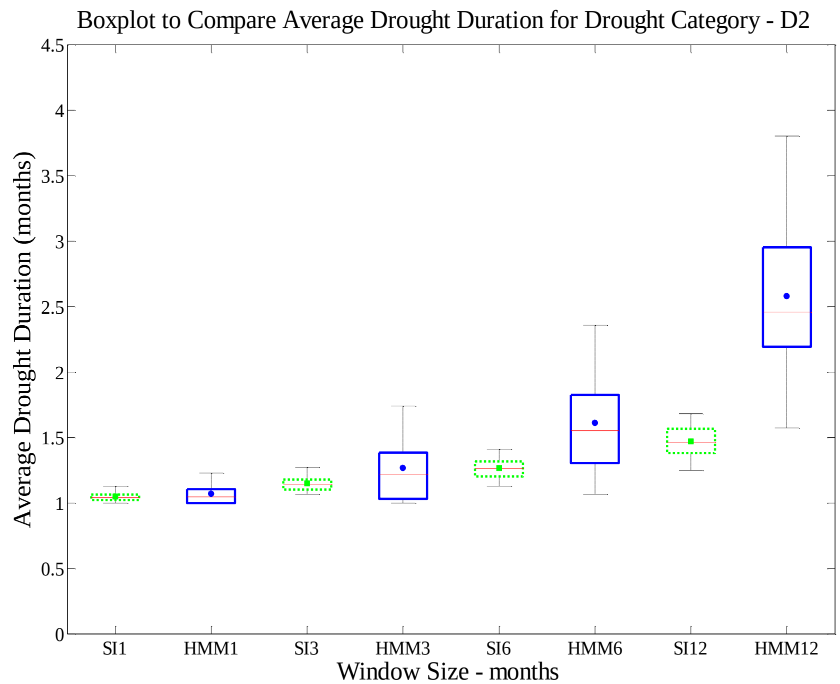


(c)



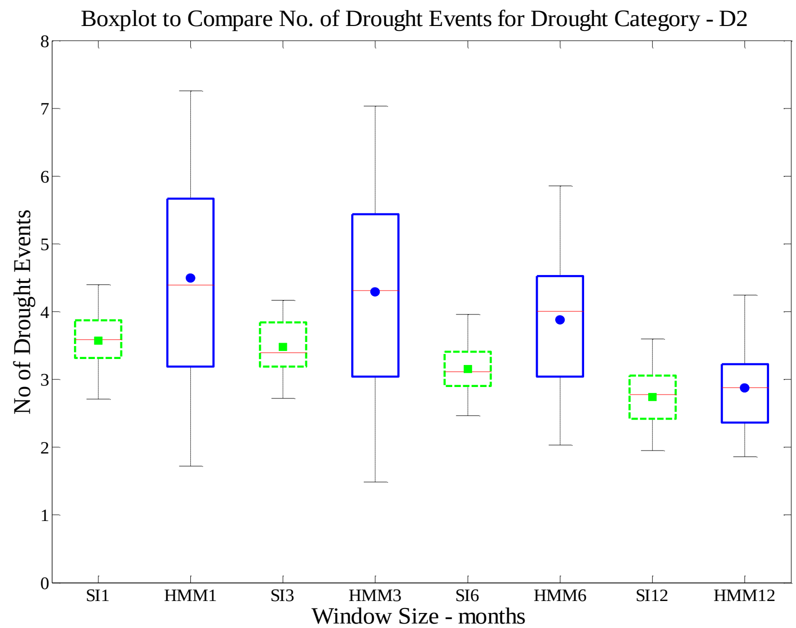
(d)

Figure 4: Graph showing HMM-DI and SSI drought classification for a 1,3,6 and 12-month(s) window at Whitewater River at Alpine,IN (3275000) station for a 5-year block 1985-89. The line plot corresponds to the cumulative streamflow total at the location used for computing the results based on the window size.



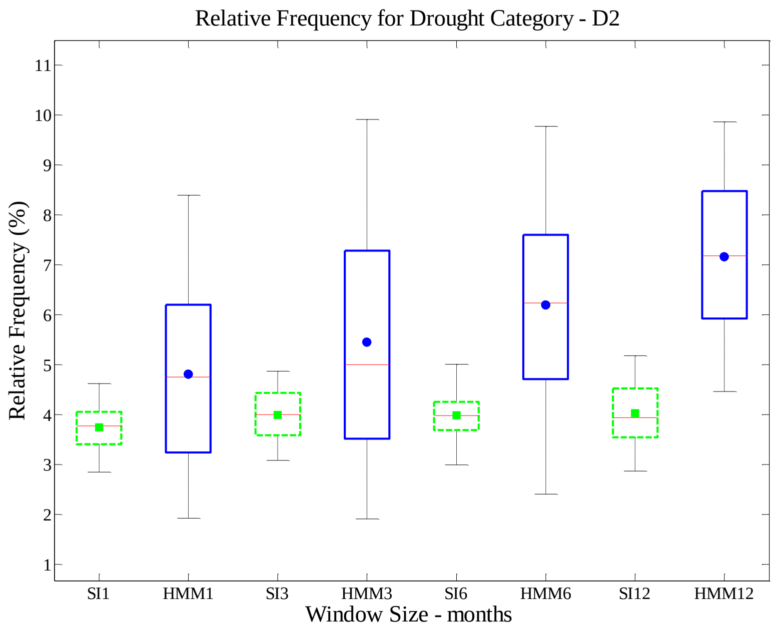
(a)

Figure 5(a): Boxplot comparing the average duration of precipitation droughts in month(s) versus window size for drought category D2. SI3 and HMM3 correspond to results from standardized index and HMM-DI for a 3 month window. The boxplot shows the variability over all 75 COOP stations over Indiana from Fig. 1.



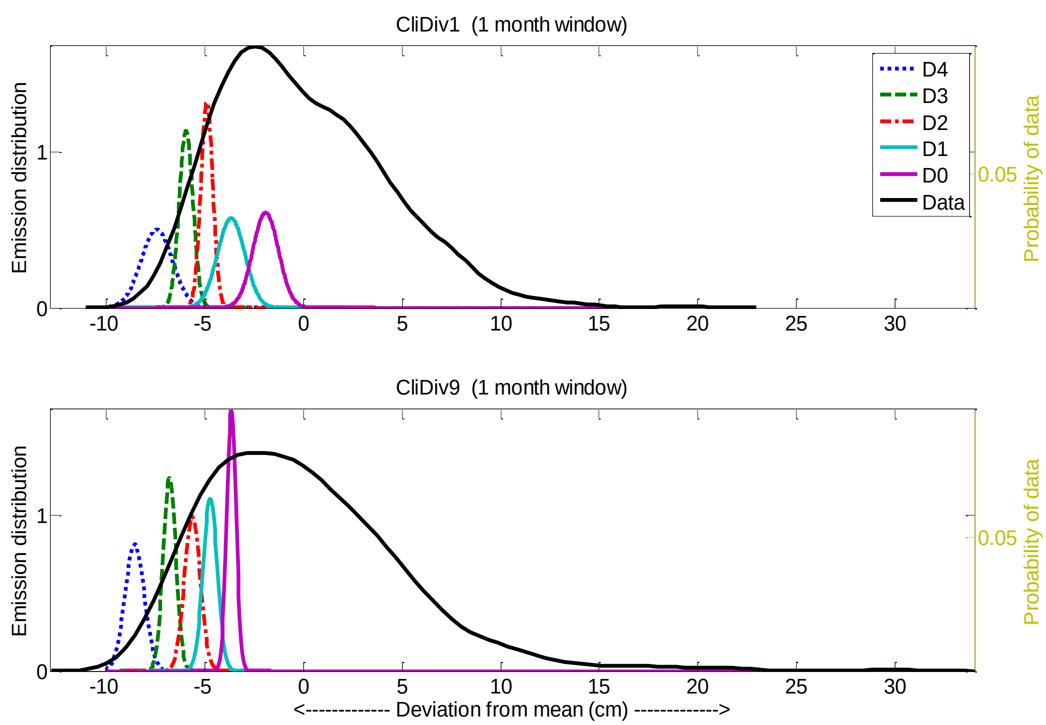
(b)

Figure 5(b): Boxplot comparing number of precipitation drought events versus window size for drought category D2 given 100 months of data. SI3 and HMM3 correspond to results from standardized index and HMM-DI for a 3 month window. The boxplot shows the variability over all 75 COOP stations over Indiana from Fig. 1.

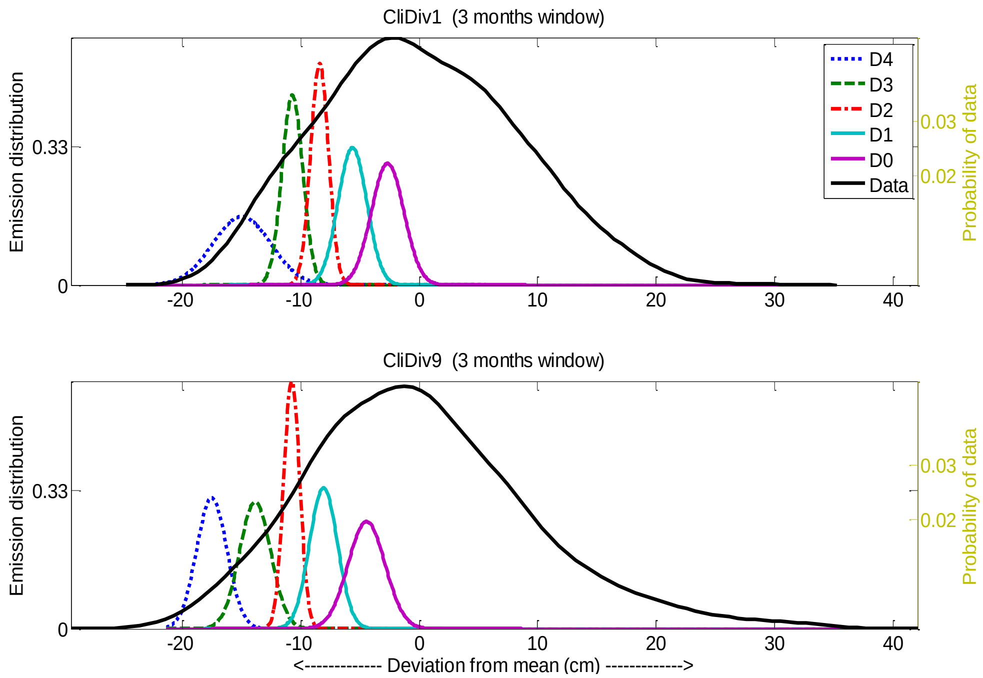


(c)

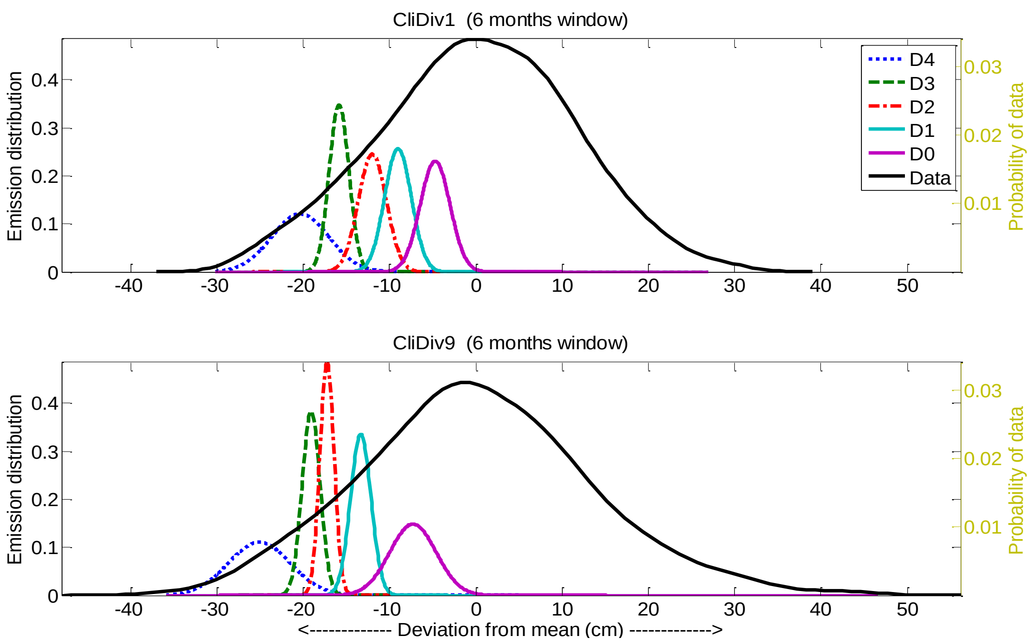
Figure 5(c): Boxplot comparing relative frequency of occurrence of D2 precipitation droughts versus window size. SI3 and HMM3 correspond to results from standardized index and HMM-DI for a 3 month window. The boxplot shows the variability over all 75 COOP stations over Indiana from Fig. 1.



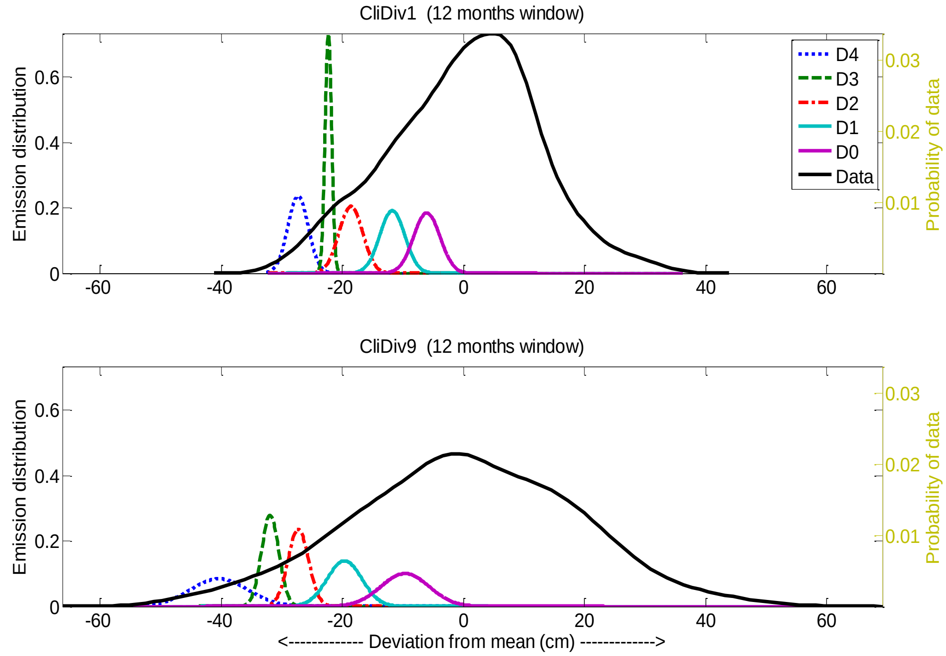
(a)


CliDiv1 (3 months window)
CliDiv9 (3 months window)
Emission distribution
Probability of data
D4
D3
D2
D1
D0
Data
<------------ Deviation from mean (cm) ------------>

(b)

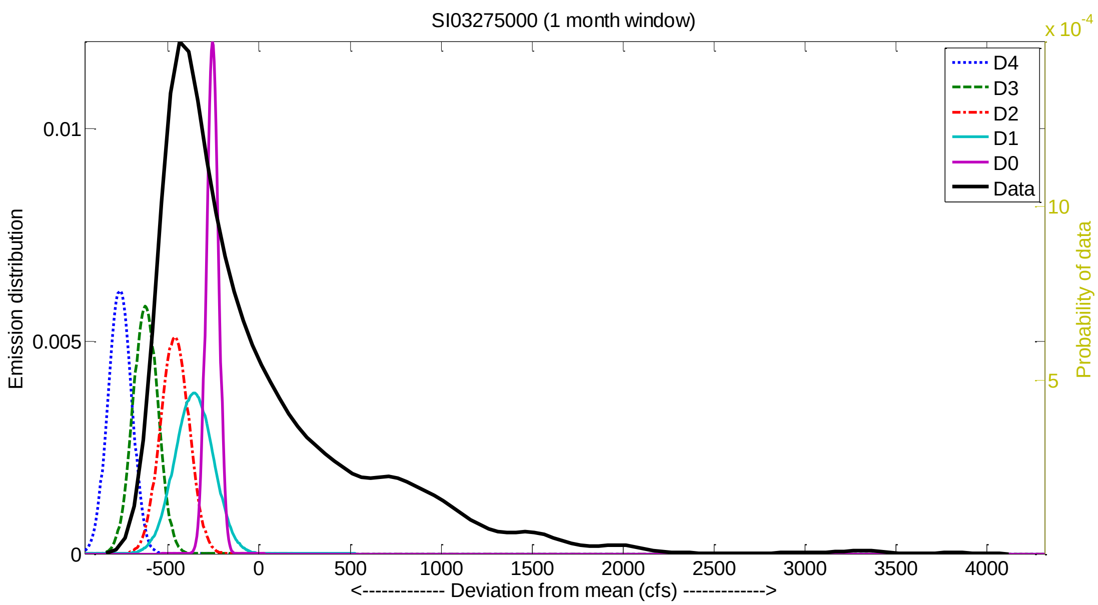

CliDiv1 (6 months window)
CliDiv9 (6 months window)
Emission distribution
Probability of data
D4
D3
D2
D1
D0
Data
<------------ Deviation from mean (cm) ------------>

(c)

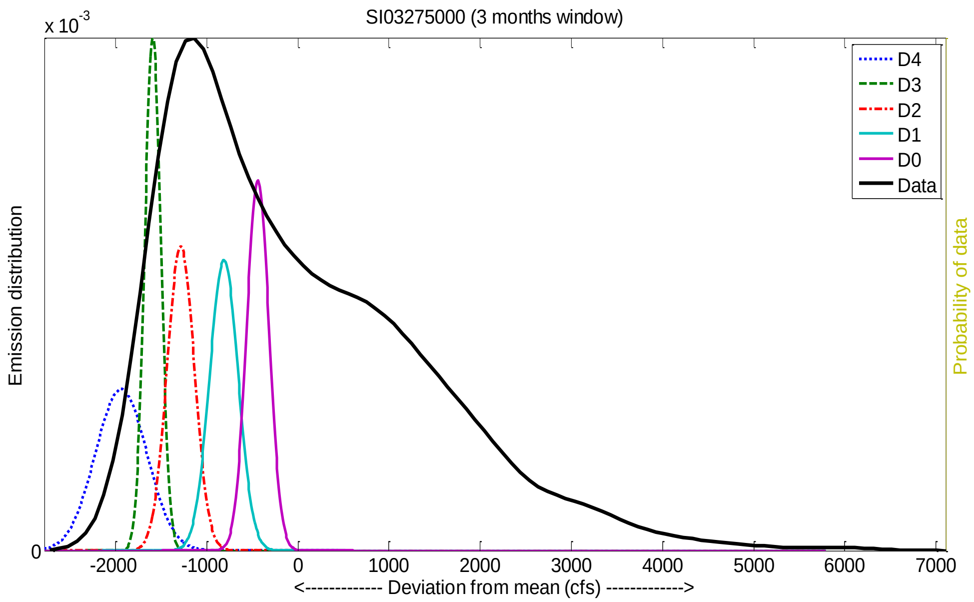


(d)

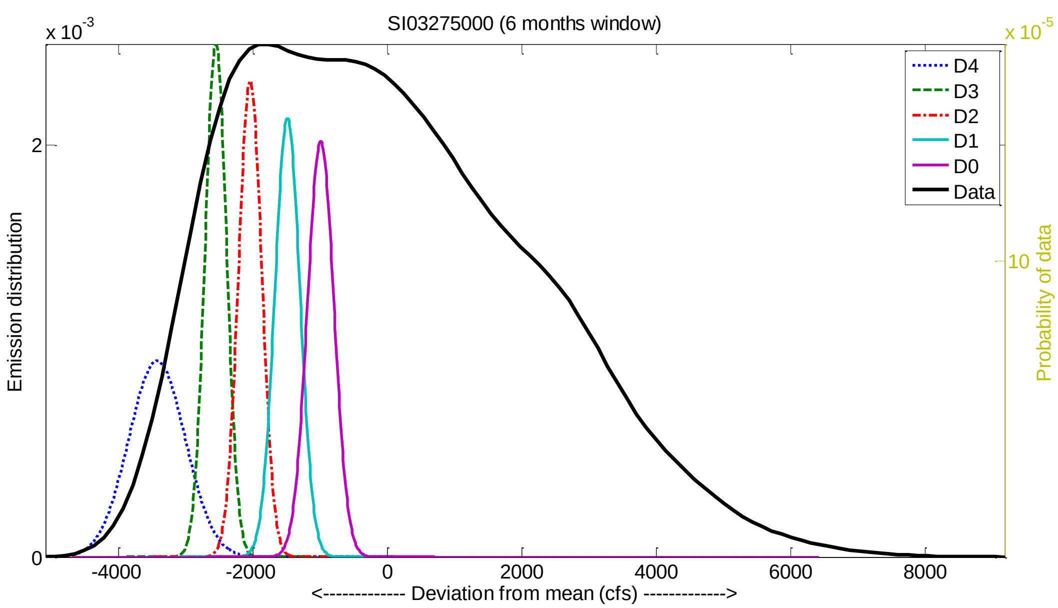
Figure 6: Plot of emission distributions for all drought states (D0-D4) estimated for climate divisions 1 and 9. The emission distributions correspond to 1, 3, 6 and 12 months time window. The thick black line represents the probability density function (pdf) of the cumulative precipitation in that window.



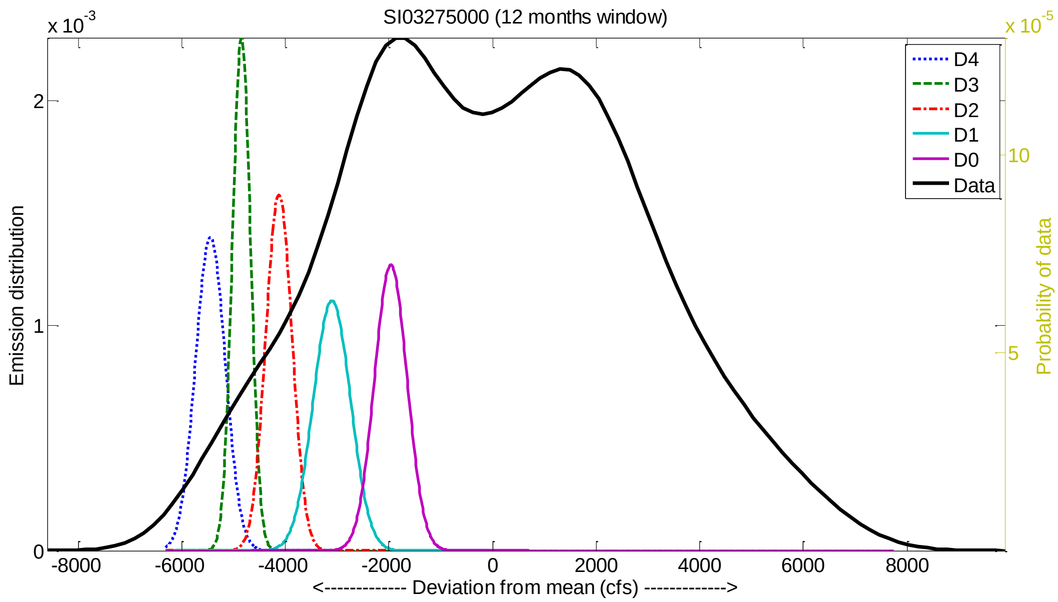
(a)


SI03275000 (3 months window)
x 10-3
Emission distribution
Probability of data
0
-2000
-1000
0
1000
2000
3000
4000
5000
6000
7000
<------------- Deviation from mean (cfs) ------------->
D4
D3
D2
D1
D0
Data

(b)


SI03275000 (6 months window)
x 10-3
x 10-5
Emission distribution
Probability of data
0
2
10
-4000
-2000
0
2000
4000
6000
8000
<------------- Deviation from mean (cfs) ------------->
D4
D3
D2
D1
D0
Data

(c)



(d)

Figure 7: Plot of emission distributions for all drought states (D0-D4) estimated at Whitewater River at Alpine,IN (3275000) USGS streamflow station. The emission distributions correspond to 1, 3, 6 and 12 month(s) time window. The thick black line represents the probability density function (pdf) of the cumulative streamflow in that window.

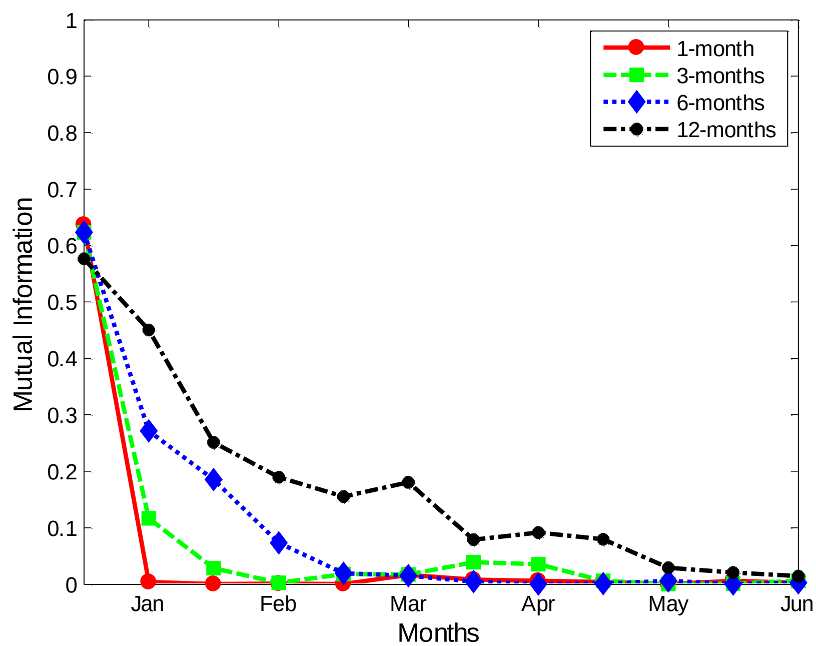


Figure 8a: Mutual information statistics for January month SPI at different time scales (1, 3, 6 and 12 months) using two bins – D4-D0 and N-W4.

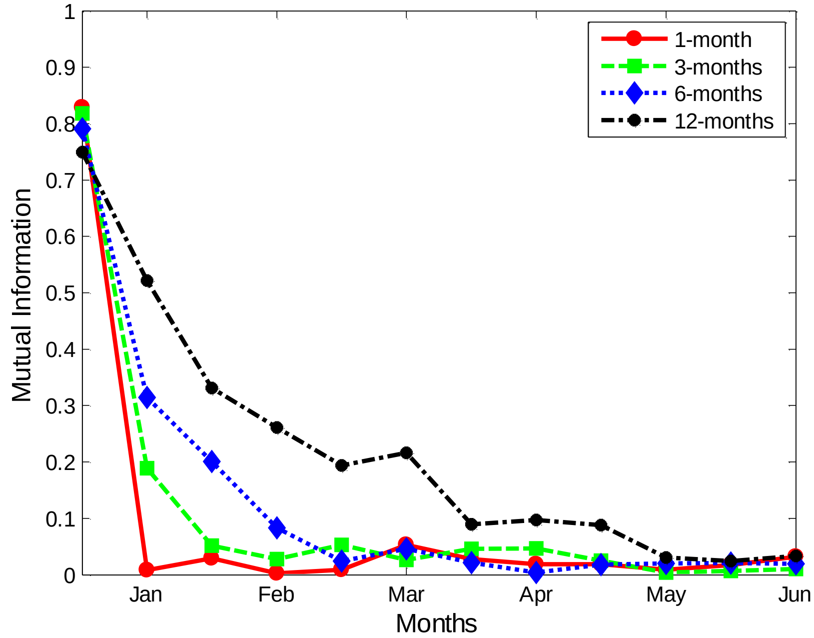


Figure 8b: Mutual information statistics for January month SPI at different time scales (1, 3, 6, and 12 months) using three bins – D4-D2, D1-D0, and N-W4.

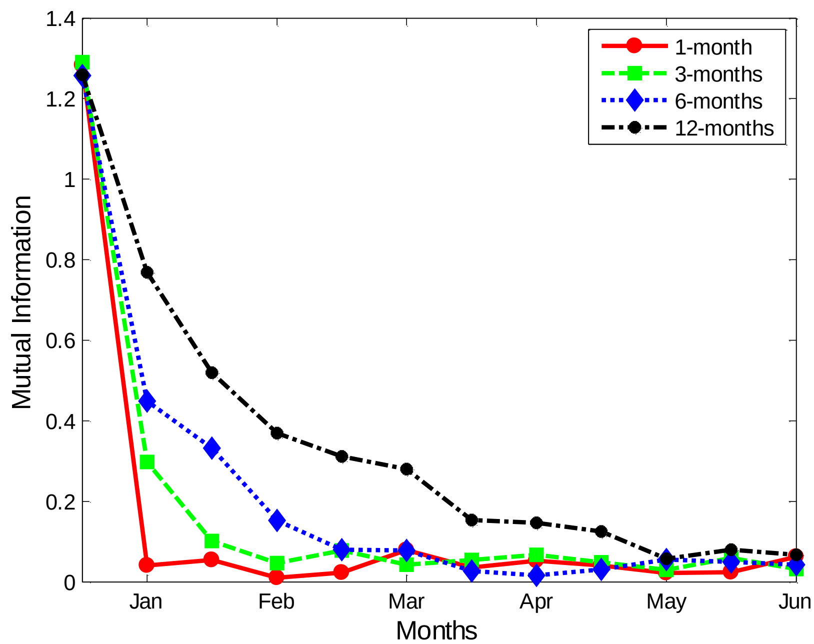


Figure 8c: Mutual information statistics for January month SPI at different time scales (1, 3, 6, and 12 months) using four bins – D4-D2, D1-D0, N, and W0-W4.