Journal of Hydrology 526 (2015) 116–126

Contents lists available at ScienceDirect

Journal of Hydrology

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

Probabilistic drought classification using gamma mixture models

Ganeshchandra Mallya [a,*], Shivam Tripathi [b], Rao S. Govindaraju [a]

[a] School of Civil Engineering, Purdue University, West Lafayette, IN 47907, USA
[b] Department of Civil Engineering, Indian Institute of Technology, Kanpur, UP 208016, India

ARTICLE INFO

Article history:
Available online 11 November 2014

Keywords:
Droughts
Standardized Precipitation Index
Probabilistic SPI
Probabilistic drought classification
Gamma mixture models
Bayesian inference

SUMMARY

Drought severity is commonly reported using drought classes obtained by assigning pre-defined thresholds on drought indices. Current drought classification methods ignore modeling uncertainties and provide discrete drought classification. However, the users of drought classification are often interested in knowing inherent uncertainties in classification so that they can make informed decisions. Recent studies have used hidden Markov models (HMM) for quantifying uncertainties in drought classification. The HMM method conceptualizes drought classes as distinct hydrological states that are not observed (hidden) but affect observed hydrological variables. The number of drought classes or hidden states in the model is pre-specified, which can sometimes result in model over-specification problem. This study proposes an alternate method for probabilistic drought classification where the number of states in the model is determined by the data. The proposed method adapts Standard Precipitation Index (SPI) methodology of drought classification by employing gamma mixture model (Gamma-MM) in a Bayesian framework. The method alleviates the problem of choosing a suitable distribution for fitting data in SPI analysis, quantifies modeling uncertainties, and propagates them for probabilistic drought classification. The method is tested on rainfall data over India. Comparison of the results with standard SPI show important differences particularly when SPI assumptions on data distribution are violated. Further, the new method is simpler and more parsimonious than HMM based drought classification method and can be a viable alternative for probabilistic drought classification.

1. Introduction

Drought classification schemes classify a drought based on its severity or intensity. Water resources planners rely on drought classification to decide drought mitigation strategies and hence weather agencies throughout the world routinely issue drought classification bulletins. For example, the US Drought Monitor releases a weekly update of drought status in U.S.A. by classifying droughts into five classes – D0–D4 with the latter representing exceptional drought. India Meteorological Department (IMD) issues drought bulletins classifying droughts into three categories, namely, mild, moderate, and severe.

The most common quantitative drought classification schemes work in two steps – first, by defining a drought index using hydro-meteorological observations and next, by categorizing droughts based on pre-defined thresholds on the index value. Examples include IMD classification that uses departure of rainfall from its long period average as a drought index, and US Drought Monitor classification that, along with other indices, uses Standardized Precipitation Index (SPI) as a drought index. Mallya et al. (2012) proposed an alternative method that does not require pre-specification of thresholds. Their method provides a probabilistic drought classification by learning thresholds from the data. Both the approaches have drawbacks arising either from the limitations of the drought index or shortcomings in the procedure for defining thresholds. The following paragraphs briefly describe some of those limitations that we have attempted to address in this work.

Drought classification schemes employ drought indices that measure degree of departure of hydro-meteorological variables, such as precipitation and streamflow, from their long-term averages. Drought indices have been used for identifying droughts and their triggers (Steinemann, 2003), assessing drought status (Kao and Govindaraju, 2010), forecasting droughts (AghaKouchak, 2014), performing drought risk analysis (Hayes et al., 2004) and studying relationship of droughts with local-scale regional hydrological variables like water quality (Sprague, 2005) and large-scale climate patterns like El Niño-Southern Oscillation (Cole and Cook, 1998; Liu and Juárez, 2001; Ryu et al., 2010). Among several drought indices proposed in the literature (Dai, 2011; Heim, 2002; Mishra and Singh, 2010), the Standardized Precipitation Index (SPI; McKee et al., 1993) is very popular because of its computational simplicity and versatility in comparing different

* Corresponding author.

http://dx.doi.org/10.1016/j.jhydrol.2014.11.008
0022-1694/

hydro-meteorological variables at different time scales. In SPI, historical observations are used to compute the probability distribution of the monthly and seasonal (4-months, 6-months, and 12-months) precipitation totals. The fitted probability distributions are then normalized using the standard inverse Gaussian function to calculate SPI values. A negative value of SPI indicates precipitation less than the median rainfall, and the magnitude of departure from zero represents the severity of a drought based on which drought classes are defined. As many drought classification schemes in the literature use SPI, they inherit its weaknesses.

Standard SPI based drought classification schemes ignore uncertainties arising from data errors, model assumptions, and parameter estimations providing discrete classification. Thus, the users are not aware of inherent uncertainties in drought classification often required for making informed decisions. Further, in the context of SPI there is an ongoing debate on the selection of the parametric distribution for fitting data. McKee et al. (1995) in their original paper on SPI recommends gamma distribution. Lloyd-Huges and Saunders (2002) found gamma distribution to be an appropriate model for Europe. Guttman (1999) suggested Pearson-III distribution as the best universal model for SPI because it provides more flexibility than the gamma distribution. Rossi and Cancelliere (2003) found normal, lognormal, and gamma distributions to be suitable for different datasets in their study. Loukas and Vasiliades (2004) investigated different theoretical distributions using Kolmogorov–Smirnov (K–S) test and Chi-squared test and found Extreme Value-I distribution to be most suitable for studying drought over Thessaly, Greece. Mishra et al. (2007) argues that different distributions may be appropriate for different drought durations (window size), and recommends K–S test for choosing an appropriate distribution. Bonaccorso et al. (2013) used Lilliefors test to choose among normal, lognormal, and gamma distributions while Russo et al. (2013) used the three parameter generalized extreme value (GEV) distribution for SPI analysis. Thus there is no consensus on the choice of distribution for SPI analysis.

Mallya et al. (2012) uses hidden Markov model (HMM) for drought classification by conceptualizing hidden states in the model to represent drought states. Their model avoided the need for specifying thresholds for drought classification and provided probabilistic drought classification by accounting model uncertainties; however, the number of hidden states (drought classes) is pre-specified. To facilitate comparison of HMM drought classification with standard methods they specified 11 hidden states. Since the number of states is imposed on the model, it is possible that for datasets with short record length the model suffers from *over-specification problem*, i.e. the model structure is more complicated than supported by the dataset. Specifically, in the HMM context, over-specification means that the number of specified hidden states are more than that needed to model the data. Over-specification can result in *parameter identification problem* leading to unreliable results.

The main objective of this paper is to propose an alternate method for probabilistic drought classification. The proposed method adapts SPI drought classification methodology by

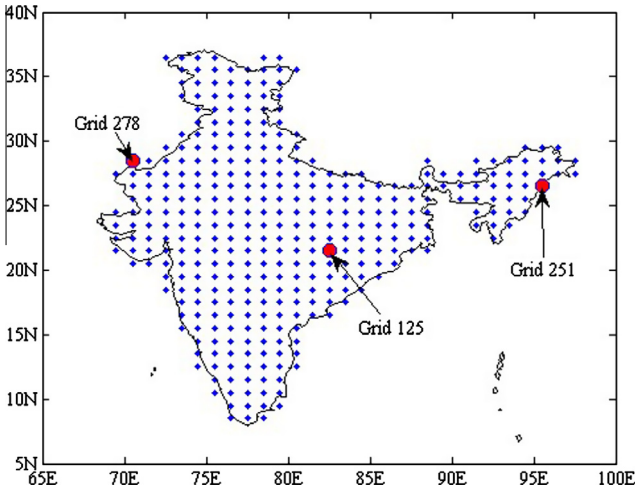

Fig. 1. Map showing the study area along with the location of grids for which rainfall data were provided by IMD.

Table 1
US Drought Monitor classification scheme. SPI ranges are prescribed for the inverse of the normal distribution. Corresponding thresholds on CDF are given in the last column.

Category	Description	SPI range	Threshold on CDF
D0	Abnormally dry	−0.5 to −0.8	0.212–0.309
D1	Moderate drought	−0.8 to −1.3	0.097–0.212
D2	Severe drought	−1.3 to −1.6	0.055–0.097
D3	Extreme drought	−1.6 to −1.9	0.023–0.055
D4	Exceptional drought	−2.0 or less	0.023 or less

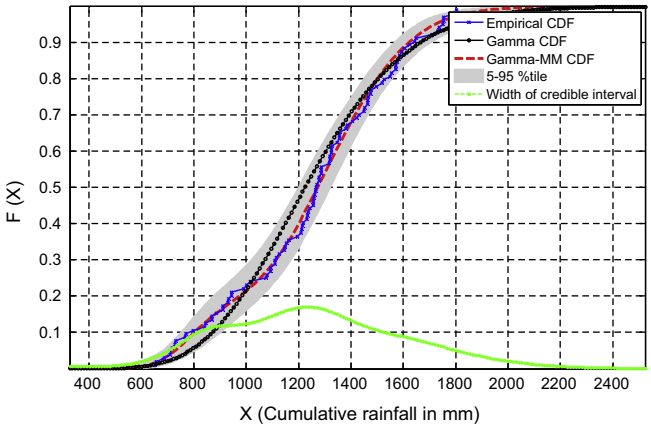

Fig. 2. Empirical CDF along with CDFs obtained by fitting gamma distribution (Gamma CDF) and gamma mixture model (Gamma-MM CDF) to the cumulative rainfall in a water-year at Grid 125. The grey band shows 5th and 95th percentile of the Gamma-MM CDF and the green dotted line shows width of its credible interval. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

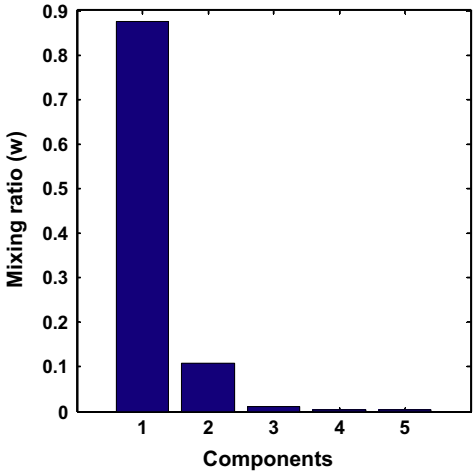

Fig. 3. Mixing ratios of the components of a Bayesian Gamma-MM. Two components are identified to be significant for characterizing water-year drought at Grid 125.

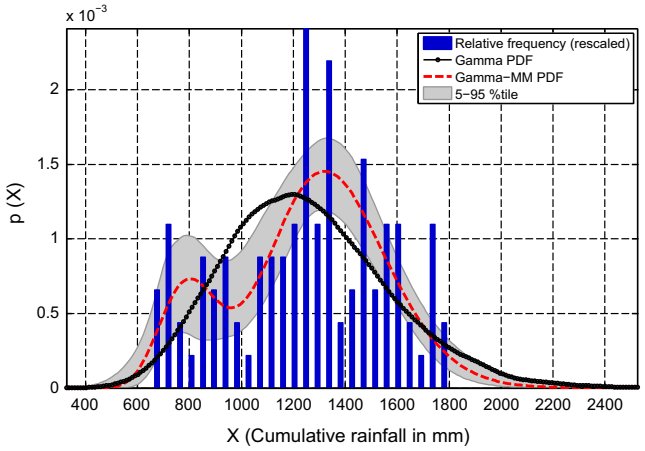

Fig. 4. Relative frequency of the cumulative rainfall amounts in a water-year at Grid 125, and probability density functions of the fitted gamma distribution (Gamma PDF) and gamma mixture model (Gamma-MM PDF). The grey band shows 90% credible interval (5th and 95th percentile) of the Gamma-MM PDF.

employing gamma mixture model (Gamma-MM) in a Bayesian framework. The method alleviates the problem of selecting suitable distribution for SPI analysis, quantifies modeling uncertainties, and propagates them for probabilistic drought classification. Further, it avoids over-specification problem by using a Bayesian approach for optimally selecting the number of hidden states in the model.

The remainder of the paper is structured as follows. First, the study area and data used are briefly described. Next, the proposed methodology for drought classification is described, and the results obtained are presented and discussed. Finally, summary and conclusions drawn from the study are presented in the last section.

2. Study area and data used

The study area, India, receives 80% of its annual precipitation during four-month long southwest summer monsoon (Bagla, 2006; Parathasarathy et al., 1994). The monsoon precipitation makes landfall around the 1st week of June near Kerala in southern India, and moves northeast towards the Himalayas. By the first week of July, almost the entire country typically receives some precipitation that continues until the end of September (Burroughs, 1999). Though the Indian monsoon is believed to be one of the most stable monsoon systems (Houghton et al., 2001), it has large inter- and intra-seasonal variability that can sometimes result in weak monsoon or droughts over India (Krishnamurthy and Shukla, 2000). Since, the country's gross domestic product (GDP), particularly food and power production, is closely linked to monsoon rains, various strategies have been developed over the years to mitigate the effects of droughts (e.g. Drought Prone Areas Programme (DPAP), and Desert Development Programme (DDP)). Effective implementation of these strategies requires real-time reliable classification of droughts.

Daily rainfall data at a spatial resolution of 1° for both latitude and longitude were obtained from India Meteorological Department (IMD) and are based on a total 1803 stations distributed over India that have at least 90% availability for the period 1901–2004 (Rajeevan, 2006). The gridded data consisting of 357 grid points have been obtained by interpolating raingage data. The IMD datasets are standard datasets widely used in monsoon-related studies over India (Goswami et al., 2006). Fig. 1 shows the study area along with the grid locations for which rainfall data were available.

3. Methodology

The proposed methodology is an adaptation of the standard SPI methodology. It classifies droughts as follows:

1. Decide a drought duration (time-window) and estimate cumulative rainfall during that period. For example, to estimate drought during a monsoon season, estimate cumulative rainfall during four months of the monsoon season (JJAS) for each year. This will yield an annual time-series of cumulative rainfall.
2. Fit a gamma mixture model (Gamma-MM) to the annual series using the procedure described in the next section. This will yield posterior distribution of model parameters.

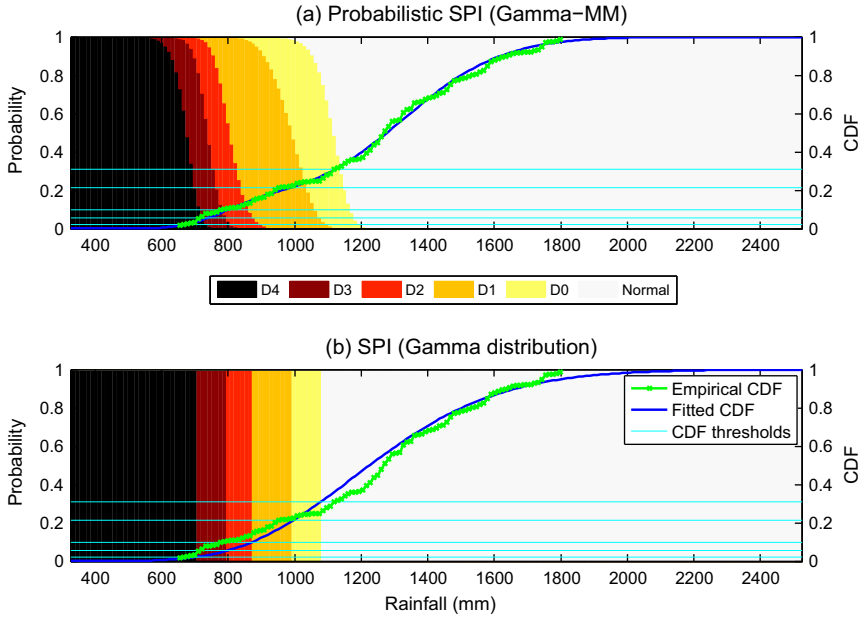

Fig. 5. Drought classification using rainfall at Grid 125 by the probabilistic SPI (top panel) and standard SPI (bottom panel). The colored patches represent drought classes, the light horizontal lines denote thresholds on CDF specified by US Drought Monitor, and the solid curves represent empirical and fitted CDFs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. For a given rainfall event, determine cumulative distribution function (CDF) and its *credible interval* using the fitted Gamma-MM. Unlike SPI, the CDF from Gamma-MM is a random variable with a distribution uniquely determined by the parameters of the fitted model.
4. Using pre-specified thresholds on the CDF, determine the drought class. As the CDF for a given rainfall event is a distribution, it may spread over more than one drought class. Estimate the mass of the CDF distribution in each drought class which will be the probability of the given rainfall event to be in that drought class.

Since the posterior distribution of the Gamma-MM parameters does not have a closed form, the integration for estimating mass of CDF in each drought class is performed numerically.

Threshold on the CDF function should be decided based on the application of the drought classification scheme. To draw parallels with the US Drought Monitor, we have used the same thresholds as used by them for SPI drought classification (Table 1).

4. Gamma mixture model (Gamma-MM)

As discussed in the Introduction section, there is an ongoing debate on the choice of a suitable distribution for fitting data in SPI analysis. We address this problem by using the gamma mixture model (Gamma-MM). Given sufficient number of components in the mixture, the Gamma-MM is proven to provide arbitrarily close approximation to any general continuous distribution in the range $(0, \infty)$ (see, DeVore and Lorentz, 1993).

The use of Gamma-MM is not new in hydrology. To model data with multiple modes and different types of skewness, (Evin et al., 2011) proposed the use of Gamma-MM for strictly positive hydrological data. In the assessment of hydrological droughts for Yellow River in China, (Shiau et al., 2007) first fitted mixtures of exponential and gamma distributions to drought duration and drought severity, respectively, and then used the copula method to construct a bivariate drought distribution. In the following we provide a brief description of the Gamma-MM. The readers are referred to Wiper et al. (2001) and Richardson and Green (1997) for details on mixture models.

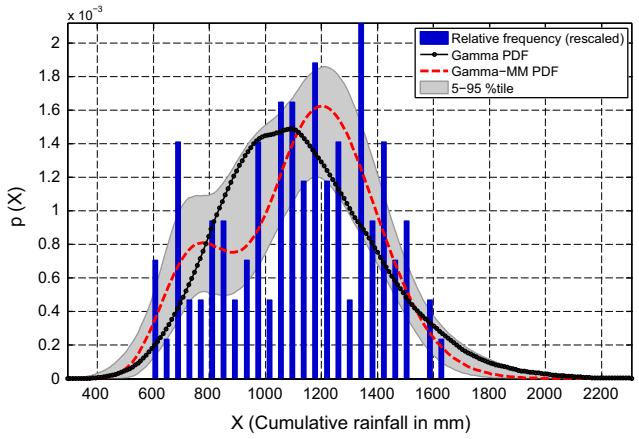

Fig. 7. Relative frequency of the cumulative rainfall amounts during the south-west summer monsoon months (JJAS) at Grid 125, and probability density functions of the fitted gamma distribution (Gamma PDF) and gamma mixture model (Gamma-MM PDF). The grey band shows 90% credible interval (5th and 95th percentile) of the Gamma-MM PDF.

Let the cumulative rainfall at time t be denoted by x_t, $t = 1, \ldots, N$ $\{x_t \in R$ and $X = [x_1, \ldots, x_N]^T\}$. If the total number of components of Gamma-MM, M, is known *a priori*, then the weighted sum of M mixtures of gamma is given by the equation

$$P(x_t|\lambda) = \sum_{i=1}^{M} w_i G\left(x_t|v_i, \frac{v_i}{\mu_i}\right), \tag{1}$$

where w_i are the mixture weights or mixing ratios, and $G\left(x_t|v_i, \frac{v_i}{\mu_i}\right)$ are the components of Gamma densities of the form,

$$G\left(x_t|v_i, \frac{v_i}{\mu_i}\right) = \frac{\left(\frac{v_i}{\mu_i}\right)^{v_i}}{\Gamma(v_i)} x_t^{(v_i-1)} \exp\left(-\frac{v_i}{\mu_i} x_t\right), \tag{2}$$

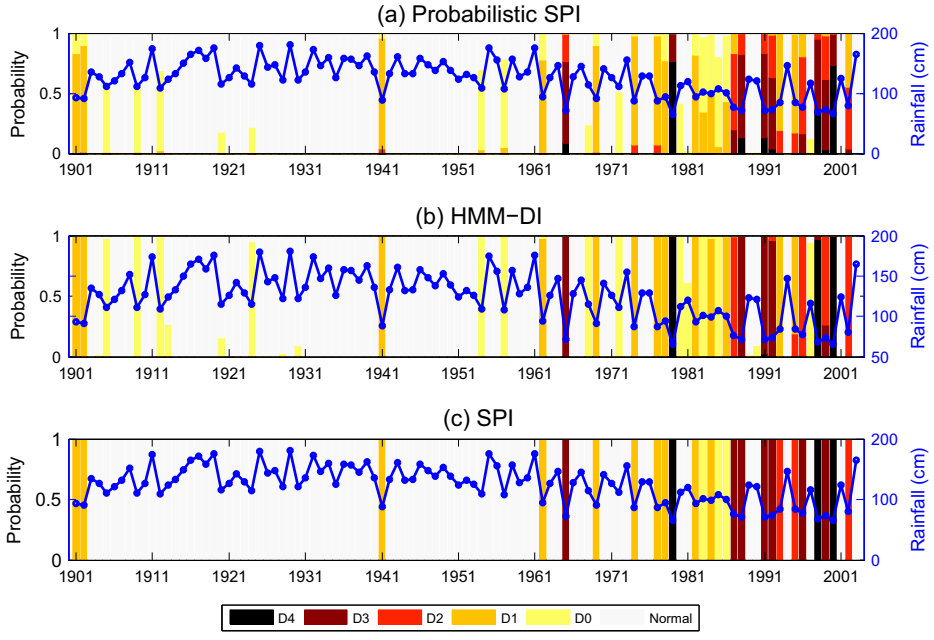

Fig. 6. Classification of historical droughts during a water-year at Grid 125 using probabilistic SPI, HMM-DI, and standard SPI approaches. The solid blue line represents cumulative rainfall during a water-year, a colored bar denotes drought classes and its length represents probability of drought state. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

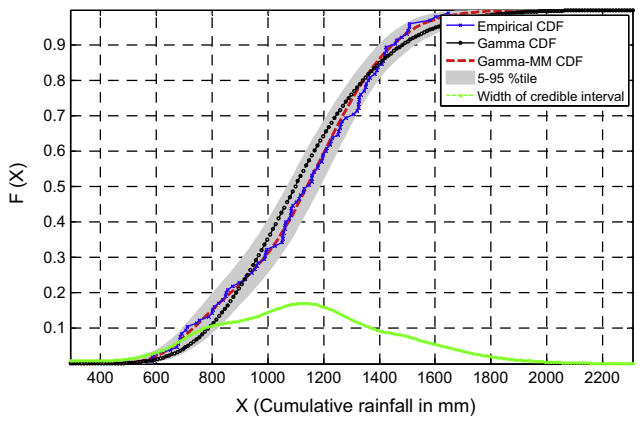

Fig. 8. Empirical CDF along with CDFs obtained by fitting gamma distribution (Gamma CDF) and gamma mixture model (Gamma-MM CDF) to the cumulative rainfall during the south-west summer monsoon months (JJAS) at Grid 125. The grey band shows 5th and 95th percentile of the Gamma-MM CDF and the green dotted line shows width of its credible interval. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

with mean μ_i and shape parameter v_i. Further, the mixture weights satisfy the constraint $\sum_{i=1}^{M} w_i = 1$. The parameter set is represented as, $\lambda = \{\mathbf{w}, \boldsymbol{\mu}, \mathbf{v}\}$ where $\boldsymbol{w} = [w_1, w_2, \ldots, w_M]^T$, $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_M]^T$ and $\boldsymbol{v} = [v_1, v_2, \ldots, v_M]^T$.

In the Bayesian framework, the model parameters are obtained by specifying prior distribution to model parameters. The parameter estimation can be simplified by introducing a latent variable $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N]^T$ for each time step. The variable $\boldsymbol{z}_t$ is an M-dimensional binary random variable, $\boldsymbol{z}_t = [z_{t1}, \ldots, z_{tM}]^T$, in which a particular element is equal to 1 and all other elements are zero, i.e. $\sum_{i=1}^{M} z_{ti} = 1$ and $z_{ti} \in \{0, 1\}$. The variable $\boldsymbol{z}_t$ denotes the component to which the data x_t belongs, and hence it is also called an *indicator variable*. The conditional distribution of x_t given $\boldsymbol{z}_t$ is

$$P(x_t|z_{ti} = 1) \sim G\left(x_t|v_i, \frac{v_i}{\mu_i}\right) \tag{3}$$

The posterior probability of the model parameters and latent variable are obtained by applying Bayes' Rule as

$$P(\lambda|X) \propto P(X|\lambda)P(\lambda) \tag{4}$$

where the parameter set λ includes the latent variable as well. The *likelihood function* given the latent variable is $P(X|\lambda) = P(X|Z, \boldsymbol{\mu}, \boldsymbol{v}) = \prod_{t=1}^{N}\prod_{i=1}^{M}\left(G\left(x_t|v_i, \frac{v_i}{\mu_i}\right)\right)^{z_{ti}}$.

Following Wiper et al. (2001) *the prior distribution* over the model parameter is given as

$P(\lambda) = P(Z|\boldsymbol{w})P(\boldsymbol{w})P(\boldsymbol{\mu})P(\boldsymbol{v})$ with

$$P(Z|\boldsymbol{w}) = \prod_{t=1}^{N}\prod_{i=1}^{M} w_i^{z_{ti}},$$

$$P(\boldsymbol{w}) = \mathrm{Dir}(\boldsymbol{w}|\Phi) = C(\Phi)\prod_{i=1}^{M} w_i^{\phi_i - 1}, \quad \Phi = [\phi_1, \ldots, \phi_M]^{\mathrm{T}},$$

$$P(\boldsymbol{v}) = \mathrm{Exp}(\boldsymbol{v}|\boldsymbol{\theta}) = \prod_{i=1}^{M} \frac{1}{\theta_i}\exp(-\theta_i v_i), \quad \boldsymbol{\theta} = [\theta_1, \ldots, \theta_M]^{\mathrm{T}}, \quad \text{and}$$

$$P(\boldsymbol{\mu}) = \mathrm{GI}(\boldsymbol{\mu}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{M} \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)}\mu_i^{-\alpha_i - 1}\exp\left(-\frac{\beta_i}{\mu_i}\right),$$

$\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_M]^{\mathrm{T}}$ and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_M]^{\mathrm{T}}$

where Dir, Exp, and GI represent Dirichlet, Exponential, and Inverted Gamma distributions respectively, and $C(\Phi)$ is a normalizing constant. The prior distribution is made non-informative by assigning following values to the hyper-parameters.

$\phi_i = 1; \theta_i = 0.01; \alpha_i = \beta_i = 1$ for $i = 1, \ldots, M$.

The posterior distribution $P(\lambda|X)$ does not have a closed form and has to be estimated by either deterministic approximation (variational Bayes methods) or stochastic approximation (MCMC; Markov chain Monte Carlo methods). In this study the posterior distribution is estimated using stochastic approximation by sampling posterior distribution with Gibbs sampler, an MCMC algorithm (Geman and Geman, 1984), the details of which are given in the appendix.

In the above formulation of Gamma-MM, we have assumed that the number of mixture components, M, is known. However, in a

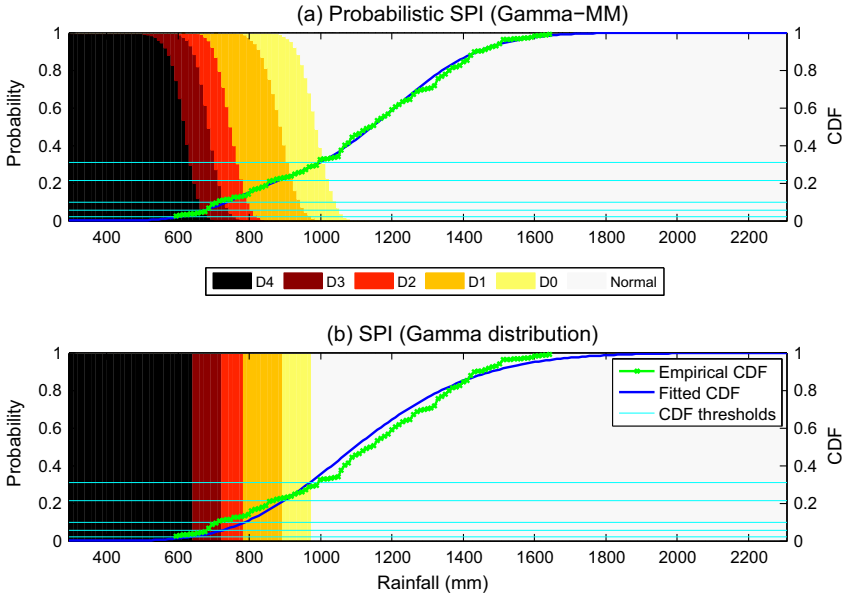

Fig. 9. Drought classification using rainfall during the south-west summer monsoon months (JJAS) at Grid 125 by the probabilistic SPI (top panel) and standard SPI (bottom panel). The colored patches represent drought classes, the light horizontal lines denote thresholds on CDF specified by US Drought Monitor, and the solid curves represent empirical and fitted CDFs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

general context, M is not known and should be estimated from data. One approach for estimating M is to consider it as a model parameter, assign prior distribution to it and estimate posterior distribution by MCMC method. Since changing M will result in a different model structure, usual MCMC algorithms such as Gibbs sampler cannot be applied. Instead reversible jump MCMC (RJMCMC; (Green, 1995) and (Richardson and Green, 1997)) may be used. In this study we implemented RJMCMC for Gamma-MM as described by Richardson and Green (1997) and Wiper et al. (2001). The results suggested that RJMCMC algorithm requires significantly higher number of iterations for convergence compared to a model where M is specified. We found that if we start with a model having sufficiently large number of components, M, the Bayesian algorithm automatically prunes the components that are not relevant for modeling by making the mixing ratio (w) very small, thereby determining optimum number of components. We recommend the latter approach for hydrological applications where the number of components is usually limited 2 or 3.

In the Bayesian framework, mixture models have *identifiability* problem i.e., a M component mixture model will have a total of $M!$ equivalent solutions. The problem can be avoided by introducing asymmetricity in the likelihood function. For example, in the context of Gamma-MM, Wiper et al. (2001) recommended the following restriction on the means of the mixture components, $\mu_1 < \mu_2 < \ldots < \mu_M$. However, for finding a good density model, as required in the present application, the problem of identifiability is not relevant because any of the equivalent solutions is as good as another (Bishop, 2006).

5. Results and discussion

The proposed approach is applied to study 4-month and 12-month droughts that correspond to a monsoon season (June to September) and water-year (June to May) drought in India, respectively. Following the procedure described in the Methodology section, first, an annual time-series of cumulative rainfall during the monsoon season and water-year are computed. Next, the droughts are classified applying the traditional SPI and the proposed approach. Both the approaches assume that cumulative time-series are stationary, and consist of independent and identically distributed samples. In the following paragraphs, results are presented for three selected grid-points (shown in Fig. 1) that reveal similarities and differences between the two drought classification approaches. As more than 80% of the rainfall in the study area is received during the monsoon season, the water-year and monsoon droughts exhibit similar characteristics. Hence, for brevity, the results are presented only at the three selected grid-points for the water-year droughts, and at only one grid point for the monsoon season. Results and discussion comparing the proposed probabilistic SPI with HMM-based probabilistic drought classification at one grid point in the study area are also included below.

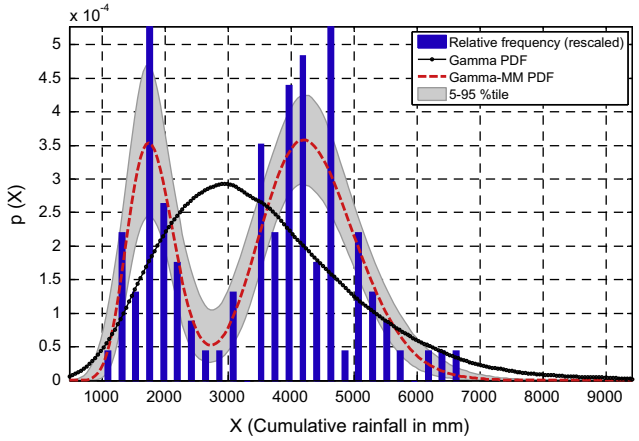

Fig. 11. Relative frequency of the cumulative rainfall amounts in a water-year at Grid 251 in NE India, and probability density functions of the fitted gamma distribution (Gamma PDF) and gamma mixture model (Gamma-MM PDF). The grey band shows 90% credible interval (5th and 95th percentile) of the Gamma-MM PDF.

a. Grid 125 (21°30′N and 82°30′E): The grid point is located in the state of Chhattisgarh and belongs to the *core-monsoon region* of India. Fig. 2 shows the empirical cumulative

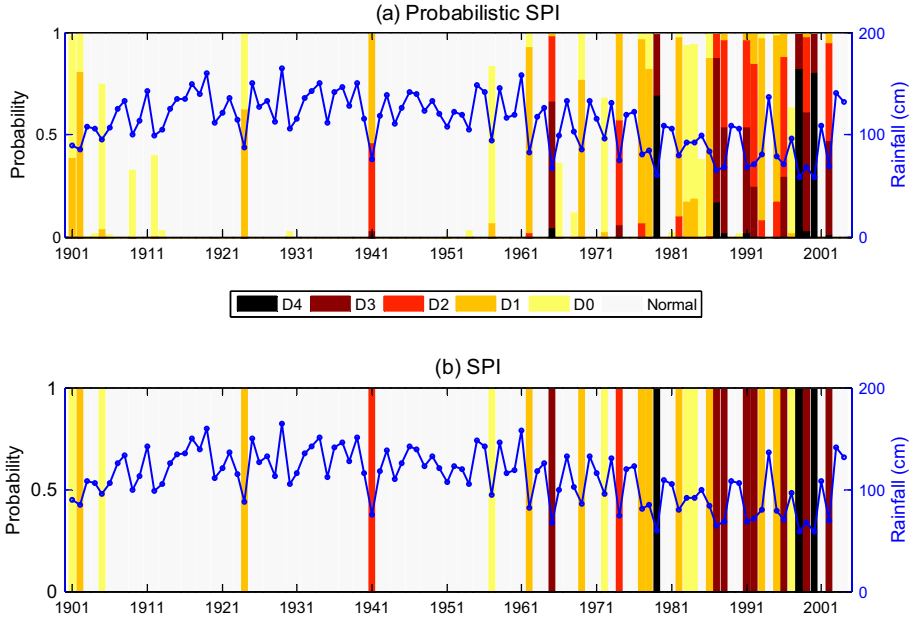

Fig. 10. Classification of historical droughts during the south-west summer monsoon months (JJAS) at Grid 125 using probabilistic and standard SPI approaches. The solid blue line represents cumulative rainfall during a water-year, a colored bar denotes drought classes and its length represents probability of drought state. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

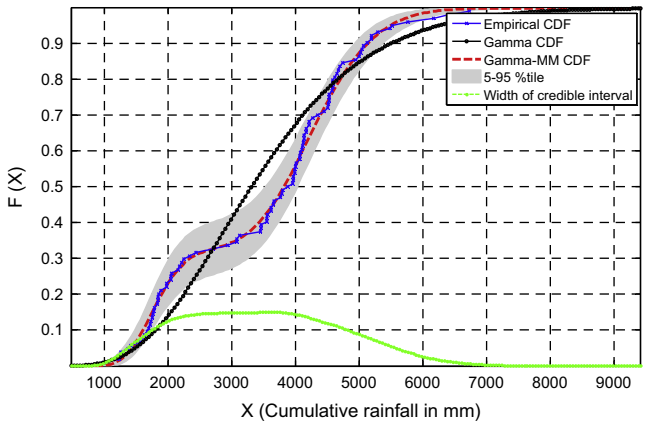

Fig. 12. Empirical CDF along with CDFs obtained by fitting gamma distribution (Gamma CDF) and gamma mixture model (Gamma-MM CDF) to the cumulative rainfall in a water-year at Grid 251 located in NE India. The grey band shows 5th and 95th percentile of the Gamma-MM CDF and the green dotted line shows width of its credible interval. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

distribution function (CDF) obtained by using Weibull plotting position formula (Chow et al., 1988) along with CDFs of fitted gamma distribution (fitted using maximum likelihood approach) and gamma mixture model (Gamma-MM) for water-year rainfall. The CDF of Gamma-MM is closer to empirical CDF than the CDF of gamma distribution, particularly for the smaller rainfall values [$F(X) < 0.25$], which are critical for drought classification. The Gamma-MM owes its better fit to the large number of tuning parameters ($3M$-1, where M is number of components in Gamma-MM) compared to two parameter gamma distribution.

Increasing the number of mixture components (M) ensures that the model provides better fit to the data. However, it may also result in over-fitting. The proposed approach addresses this problem by using a Bayesian framework that avoids overfitting by marginalizing over the model parameters instead of making point estimates. Fig. 3 shows the mixing ratio of a 5-component Gamma-MM fitted to cumulative water-year rainfall at Grid 125. The model identifies that three of the five components have negligible contribution and are effectively pruned from the model. Thus, the Bayesian framework identifies optimal number of mixture components needed to fit the data.

The Bayesian framework also allows quantification of model uncertainties and their propagation to model estimates. In the context of Gamma-MM, the posterior distribution of model parameters is estimated from which the CDF is obtained. Unlike maximum likelihood approach that yields a point estimate of CDF, the Bayesian approach treats CDF as a random variable and yields distribution of CDFs for a given value of rainfall. The grey shaded band in Fig. 2 represents 90% credible interval (5th and 95th percentile). The width of the credible interval is not constant but varies with the magnitude of rainfall. It has a maximum value of 0.16 near the median rainfall (1260 mm), a plateau near the intersection of two components (~900 mm; Fig. 4), and a monotonic decreasing trend on either side of the median.

The width of the credible interval is large even for smaller values of CDFs that decide drought classes in SPI methodology. In this study, we attempted to engage credible interval of CDF for drought classification. Fig. 5(b) shows the drought classification using standard SPI method. The empirical CDF along with the fitted CDF and drought classification thresholds are shown in the figure. The SPI drought classification uses fixed thresholds, hence the boundaries separating two drought classes are vertical lines on the panel. The top panel of Fig. 5 shows probabilistic drought classification by using Gamma-MM. The classification uses the same thresholds on CDF as SPI but engages uncertainty in the estimate of CDF resulting in probabilistic drought classification. Unlike standard SPI, the demarcating boundaries in the probabilistic SPI are curves denoting varying classification probabilities.

The probabilities associated with drought classification represent uncertainties in determining drought classes. For example, the D4 category drought represents drought conditions

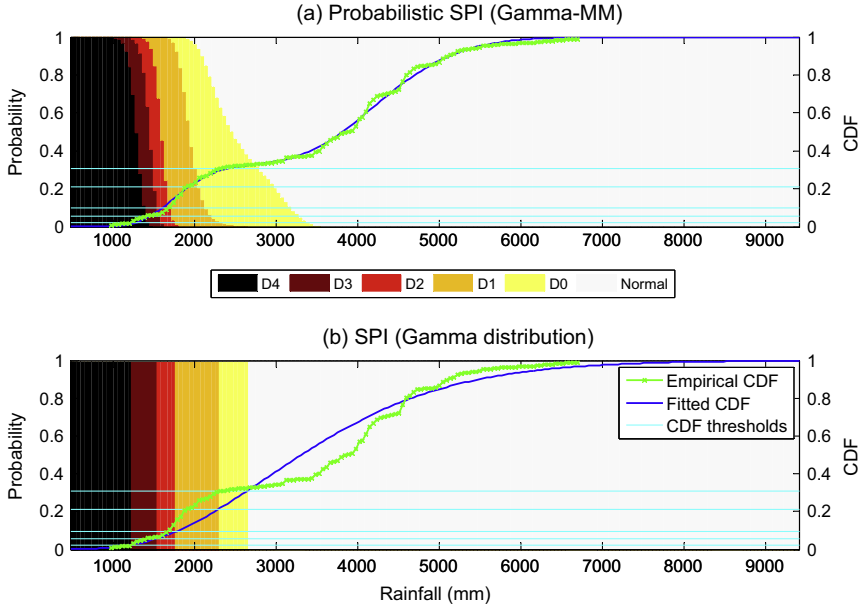

Fig. 13. Drought classification using rainfall at Grid 251 in NE India by the probabilistic SPI (top panel) and standard SPI (bottom panel). The colored patches represent drought classes, the light horizontal lines denote thresholds on CDF specified by US Drought Monitor, and the solid curves represent empirical and fitted CDFs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

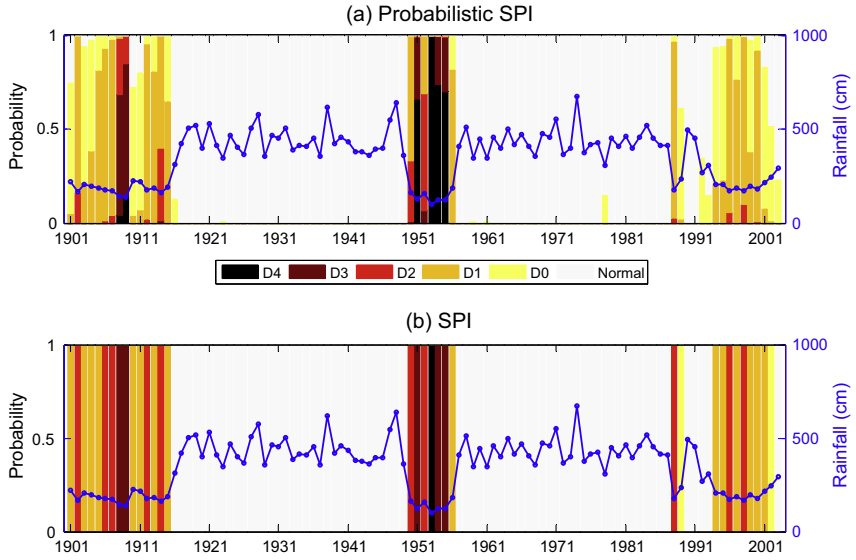

Fig. 14. Classification of historical droughts during a water-year at Grid 251 in NE India using probabilistic and standard SPI approaches. The solid blue line represents cumulative rainfall during a water-year, a colored bar denotes drought classes and its length represents probability of drought state. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

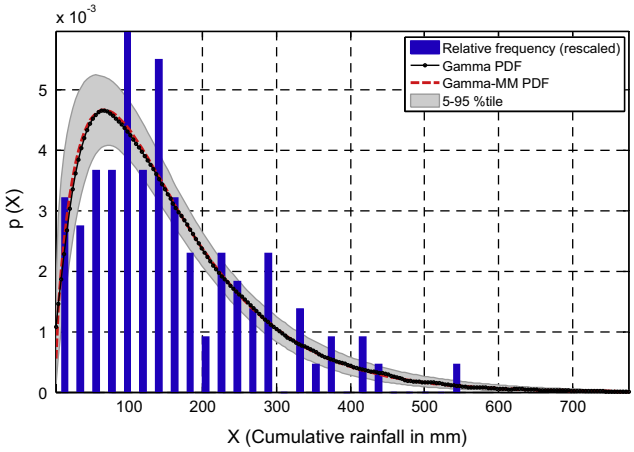

Fig. 15. Same as Fig. 11 but for Grid 278 in the Thar Desert of Western India.

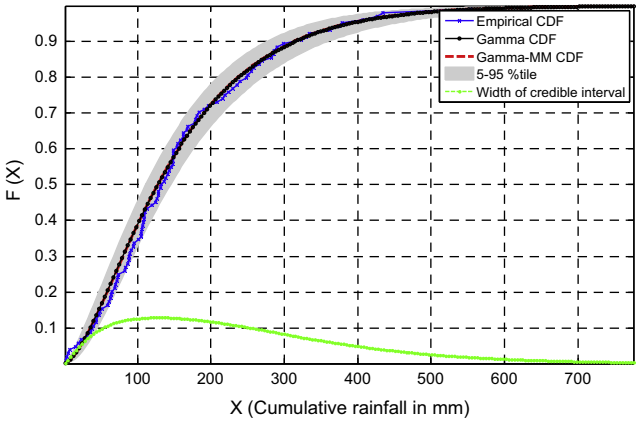

Fig. 16. Same as Fig. 12 but for Grid 278 in the Thar Desert of Western India.

where non-exceedance probability of the cumulative rainfall is less than 0.023 [$F(X) < 0.023$; Table 1], i.e. during D4 drought the chance of rainfall being less than the observed rainfall is less than 2%. The probabilistic drought classification acknowledges that, given limited data and model assumptions, such a threshold cannot be determined uniquely but can be estimated probabilistically. The method honors model uncertainty and provides results in a format that could be useful for drought managers.

Fig. 6 shows historical drought classes at Grid 125 using standard SPI, probabilistic SPI, and HMM based drought classification (HMM-DI). The droughts classified by probabilistic SPI (Fig. 6a) and standard SPI methods (Fig. 6c) are similar, however, the advantages of probabilistic classification are evident in some years. For example, in 1998, 1999 and 2000 the cumulative rainfall values were 69 cm, 73 cm, and 66 cm, respectively. Considering that the difference in cumulative rainfall among these years is less than 3% of their standard deviation (30 cm), we would not have expected them to belong to two different drought classes as categorized by SPI (1998 and 2000 in D4, and 1999 in D3). The probabilistic SPI classifies 1998, 1999 and 2000 to D3 class with probability 55%, 60% and 25%, and to D4 class with probabilities 40%, 5% and 75%, respectively (the remaining probabilities being given to other drought classes). The historical drought classes at Grid 125 using HMM-DI are shown in Fig. 6b. Compared to the probabilistic SPI results (Fig. 6a), drought classes obtained using HMM are more conservative. This is evident for the years 1920, 1924, 1998 and 2000 where droughts are classified with higher probabilities, or in a more severe category by HMM-DI compared to drought classification using probabilistic SPI. An HMM with 11 hidden states may suffer from an over-specification problem.

Fig. 7 shows the relative frequency of the rainfall during the monsoon months, JJAS, at Grid 125. As in the case of water-year rainfall (Fig. 4), the monsoon rainfall also exhibits two distinct modes that are captured by the 2-component Gamma-MM but missed by the gamma distribution. Fig. 8 shows the empirical CDF of the monsoon rainfall along with CDFs of the fitted gamma distribution, and Gamma-MM model with its 90% credible interval. The width of the credible interval is widest (0.17) near the median rainfall (1140 mm), a plateau at the intersection of two components of the Gamma-MM (~800 mm, Fig. 7) and a monotonic decreasing trend away

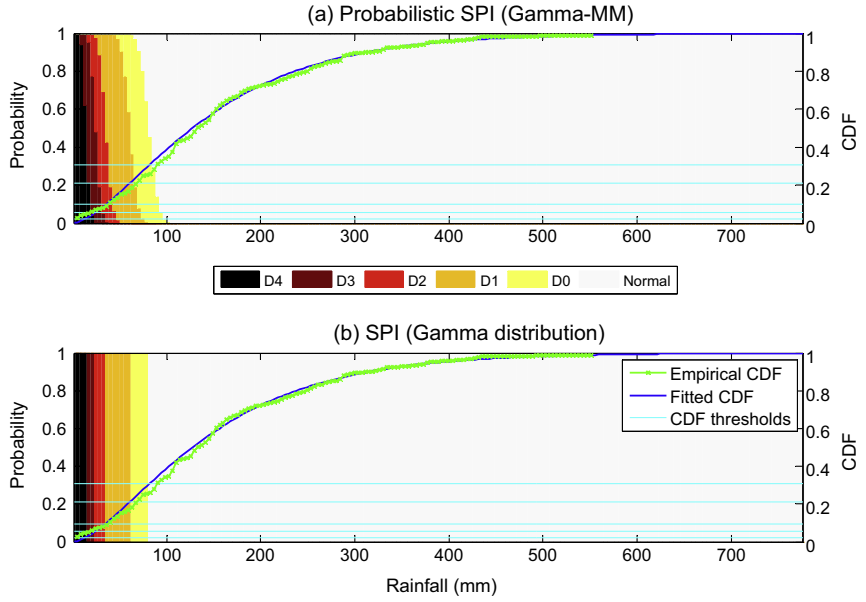

Fig. 17. Same as Fig. 13 but for Grid 278 in the Thar Desert of Western India.

from the median, similar in nature to Fig. 4. Fig. 9 presents the demarcating boundaries for the drought classes determined by the two methods. As in the case of water-year droughts (Fig 5), the demarcating boundaries for probabilistic SPI are S-shaped curves. The classification of historical monsoon droughts by standard SPI and, probabilistic SPI are similar except for some subtle differences (Fig. 10). In 1901, 1902 and 1924 the monsoon rainfall at Grid 125 were 90 cm, 85 cm and 88 cm, respectively. Standard SPI classifies 1901 in D0 class, but 1902 and 1924 in D1 class even though their differences from 1901 rainfall are not significant (5 cm and 2 cm, respectively). Probabilistic SPI classifies all the three years in D0 and D1 classes with probabilities 60% & 39%, 19% & 81%, and 37% & 63%, respectively.

b. Grid 251 (26°30′N and 95°30′E): The grid is located in North-East India, which is among the highest rainfall receiving regions of the world. Fig. 11 shows the relative frequency of the rainfall received during a water year. The data exhibits two distinct modes that are captured by the 2-component Gamma-MM but completely missed by the gamma distribution. Fig. 12 shows the empirical CDF of the cumulative rainfall along with CDFs of the fitted gamma distribution, and Gamma-MM model with its 90% credible interval. The credible interval is widest near the intersection of two components of the Gamma-MM (Fig. 7). Fig. 13 presents the demarcating boundaries for the drought classes determined by the two methods. A notable feature in the figure is a relatively diffused boundary separating D0 category drought from the normal state in probabilistic SPI which can be attributed to a relatively wide credible interval in that range (2500–3500 mm, Fig. 12). The drought classification of the historical data is given in Fig. 14. Compared to standard SPI, the probabilistic SPI is more conservative in assigning D4 category drought. For example, 1953, 1954 and 1955

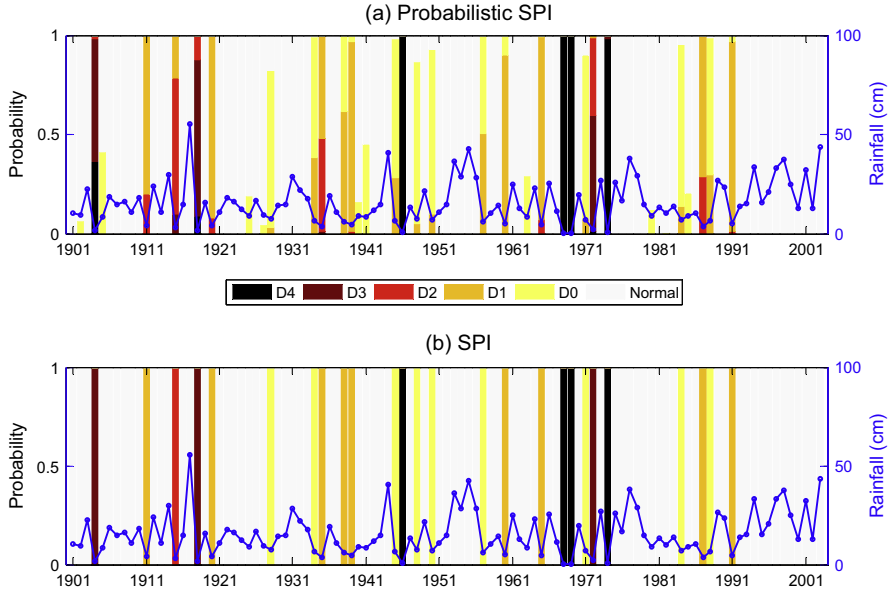

Fig. 18. Same as Fig. 14 but for Grid 278 in the Thar Desert of Western India.

are the lowest rainfall years in the record with cumulative rainfall of 98 cm, 124 cm and 125 cm, respectively. Standard SPI classifies only 1953 in D4 class while probabilistic SPI classifies all the three years in D4 class with probabilities 99%, 74% and 71%, respectively.

c. Grid 278 (28°30′N and 70°30′E): The grid belongs to the Thar Desert in western India where the annual rainfall is much smaller than rest of the country. Fig. 15 shows the relative frequency of the cumulative rainfall during a water year along with PDFs of gamma distribution and Gamma-MM. The Gamma-MM selects only one component and yields a distribution that is very similar to that of gamma distribution (Figs. 15 and 16). The 90% credible interval shows a peak near 100 cm which lies in the tail of the rainfall distribution and has implications on drought classification. Fig. 17 illustrates drought classification by the standard SPI and probabilistic SPI. The two methods provide similar drought classification except for a few minor differences. The cumulative rainfall of 100 cm represents normal state according to standard SPI classification, however owing to wide credible interval, the rainfall is assigned to D0 drought category by probabilistic SPI, albeit with a small probability (1.5%). The classifications of the historical droughts by the two methods are almost similar (Fig. 18). Thus, for the scenarios where data support the gamma distribution assumption of SPI, the results of Gamma-MM based probabilistic SPI and standard SPI are similar.

6. Summary and concluding remarks

1. A probabilistic drought classification method is proposed as an alternative to (i) deterministic classification by standard SPI, and (ii) probabilistic classification by HMM.
2. The proposed method alleviates the problem of choosing a suitable distribution for SPI analysis by modeling the data with a mixture of gamma distributions. Given sufficient components in the mixture, the Gamma-MM can give arbitrarily close approximation to any general continuous distribution in the range $(0, \infty)$.
3. The problem of overfitting the data is avoided by using Bayesian framework that determines optimum number of components needed by the model.
4. The proposed method propagates model uncertainties to drought classification by providing probabilistic drought classes.
5. The method was tested on rainfall data over India. Specifically, droughts during the water year (June–May) and the south-west monsoon season (JJAS) were studied in detail using the proposed method. The results suggest that drought classification by the proposed method is similar to standard SPI classification where data satisfies SPI assumptions. However, the results of the new method are markedly different and more intuitive than SPI results for situations where data violate SPI assumptions. The drought classification obtained using the proposed method were less conservative compared to the probabilistic classification by HMM with 11 hidden states as it avoids the problem of over-specification.

The proposed Gamma-MM method for probabilistic drought classification has a slightly more involved algorithm than standard SPI, but the former quantifies uncertainty in drought classification, a critical input for hydrological decision-making (Pappenberger and Beven, 2006). Recent studies have highlighted the need of probabilistic analysis for characterizing droughts (Mishra et al., 2009), forecasting droughts (Madadgar and Moradkhani, 2013 and AghaKouchak, 2014), performing drought risk analysis (Hayes et al., 2004), determining drought recovery (Pan et al., 2013), and managing droughts (Song, 2011). The proposed approach, owing to its probabilistic framework and relatively simple algorithm compared to HMM-DI, can be a viable tool for these analyses.

In the paper, the probabilistic SPI is applied to the rainfall data. However, the proposed method can be easily extended for classifying droughts using other hydro-meteorological variables such as streamflow, runoff, groundwater, and soil moisture for which SPI like indices have been proposed in the literature. Many of these hydro-meteorological variables have large measurement uncertainties, which are ignored in standard SPI type analysis, but can be easily engaged in the proposed method. Further, the method opens avenues for defining droughts for non-stationary hydrological records and characterizing droughts in real time and using online Bayesian updates.

Acknowledgments

Studies of the authors were supported in part by the National Science Foundation under Grant DBI 0619086. 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 A. The Gibbs sampling algorithm

The Gibbs sampling algorithm samples posterior distribution of the parameters by sequentially sampling from the conditional distribution of a parameter given all other parameters. The sampling starts with an initial value and proceeds as follow.

1. Set iteration number $j = 0$, and parameters to their initial value $\boldsymbol{\lambda}^{(0)} = [\boldsymbol{w}^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{v}^{(0)}]$. The initial value is obtained by randomly sampling from the prior distribution of the parameters.
2. Sample from $P(\boldsymbol{z}_t^{(j+1)}|X, \boldsymbol{w}^{(j)}, \boldsymbol{\mu}^{(j)}, \boldsymbol{v}^{(j)}) \sim \text{Multinomial}(\boldsymbol{z}_t|\boldsymbol{r}_t)$ where $\boldsymbol{r}_t = [r_{t1}, \ldots, r_{tM}]^{\mathrm{T}}$, $r_{ti} = \frac{s_{ti}}{\sum_{i=1}^{M} s_{ti}}$ and $s_{ti} = w_i G(x_t|v_i, \frac{v_i}{\mu_i})$ and Multinomial represents multinomial distribution.
3. Sample from $P(\boldsymbol{w}^{(j+1)}|X, Z^{(j+1)}, \boldsymbol{\mu}^{(j)}, \boldsymbol{v}^{(j)}) \sim \text{Dir}(\boldsymbol{w}|\hat{\Phi})$ where $\hat{\Phi} = [\phi_i + n_i, \ldots, \phi_M + n_M]^{\mathrm{T}}$ and $n_i = \sum_{t=1}^{N} z_{ti}$.
4. Sample from $P(\boldsymbol{\mu}^{(j+1)}|X, Z^{(j+1)}, \boldsymbol{w}^{(j+1)}, \boldsymbol{v}^{(j)}) \sim \text{GI}(\boldsymbol{\mu}|\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ where $\hat{\boldsymbol{\alpha}} = [\alpha_i + n_i v_i, \ldots, \alpha_M + n_M v_M]^{\mathrm{T}}$ and $\hat{\boldsymbol{\beta}} = [\beta_i + v_i \sum_{t=1}^{M} x_t z_{ti}, \ldots, \beta_M + v_M \sum_{t=1}^{M} x_t z_{tM}]^{\mathrm{T}}$.
5. Sample from $P(\boldsymbol{v}^{(j+1)}|X, Z^{(j+1)}, \boldsymbol{w}^{(j+1)}, \boldsymbol{\mu}^{(j+1)})$. This conditional distribution does not have a closed form. Hence samples are generated using Metropolis–Hasting algorithm. In the Metropolis–Hasting algorithm a sample is generated from a proposal distribution $P(\tilde{v}_i|v_i) \sim G(h, h|v_i)$ and is accepted with a probability $\min\{1, \frac{f(\tilde{v}_i)P(v_i|\tilde{v}_i)}{f(v_i)P(\tilde{v}_i|v_i)}\}$ where $f(v_i) \propto \frac{v_i^{n_i v_i}}{\Gamma(v_i)^{n_i}} \exp\left(-v_i\left(\theta_i + \frac{\sum_t x_t z_{ti}}{\mu_i} + n_i \log \mu_i - \log\left(\prod_{t=1;z_{ti}=1}^{N} x_t\right)\right)\right)$. If the new sample $\tilde{v}_i$ is rejected, the current value of v_i is retained. The above procedure is repeated to sample v_i for all components $i = 1, \ldots, M$. In this study the parameter of the proposal distribution, h, is set to 2.
6. Set $j = j + 1$ and go to Step 2 until convergence. In this study, 15,000 samples are generated after ignoring initial 500 samples (*burn-in* period). Trace plots of the samples are monitored for convergence.

To keep the notations uncluttered, the iteration number is omitted from the parameters of the conditional distributions.

References

Bagla, P., 2006. Controversial rivers project aims to turn India's fierce monsoon into a friend. Science 313 (5790), 1036–1037. http://dx.doi.org/10.1126/science.313.5790.1036.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, New York.

Bonaccorso, B., Peres, D.J., Cancelliere, A., Rossi, G., 2013. Large scale probabilistic drought characterization over Europe. Water Resour. Manage. 27 (6), 1675–1692. http://dx.doi.org/10.1007/s11269-012-0177-z.

Burroughs, W.J., 1999. The Climate Revealed. Cambridge University Press.

Chow, V.T., Maidment, D.R., Mays, L.W., 1988. Applied Hydrology. McGraw-Hill Series in Water Resources and Environmental Engineering.

Cole, J.E., Cook, E.R., 1998. The changing relationship between ENSO variability and moisture balance in the continental United States. Geophys. Res. Lett. 25 (24), 4529–4532. http://dx.doi.org/10.1029/1998GL900145.

Dai, A., 2011. Drought under global warming: a review. Wiley Interdiscip. Rev.: Climate Change 2 (1), 45–65. http://dx.doi.org/10.1002/wcc.81.

DeVore, R.A., Lorentz, G.G., 1993. Constructive Approximation. Springer.

Evin, G., Merleau, J., Perreault, L., 2011. Two-component mixtures of normal, gamma, and Gumbel distributions for hydrological applications (W08525). Water Resour. Res. 47 (8). http://dx.doi.org/10.1029/2010WR010266.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6(6), pp. 721–741, doi: 10.1109/TPAMI.1984.4767596.

Goswami, B.N., Venugopal, V., Sengupta, D., Madhusoodanan, M.S., Xavier, P.K., 2006. Increasing trend of extreme rain events over India in a warming environment. Science 314 (5804), 1442–1445. http://dx.doi.org/10.1126/science.1132027.

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82 (4), 711–732. http://dx.doi.org/10.1093/biomet/82.4.711.

Guttman, N.B., 1999. Accepting the standardized precipitation index: a calculation algorithm. JAWRA J. Am. Water Resour. Assoc. 35 (2), 311–322. http://dx.doi.org/10.1111/j.1752-1688.1999.tb03592.x.

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

Heim, R.R., 2002. A review of twentieth-century drought indices used in the United States. Bull. Am. Meteorol. Soc. 83 (8), 1149.

Houghton, J.T., Ding, Y., Griggs, D.J., Noguer, M., Linden, P.J. van der, Dai, X., Maskell, K., Johnson, C.A. (Eds.), 2001. Climate Change 2001: The Scientific Basis. Cambridge University Press.

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

Krishnamurthy, V., Shukla, J., 2000. Intraseasonal and interannual variability of rainfall over India. J. Climate 13 (24), 4366–4377. http://dx.doi.org/10.1175/1520-0442(2000) 013<0001:IAIVOR>2.0.CO;2.

Liu, W.T., Juárez, R.I.N., 2001. ENSO drought onset prediction in northeast Brazil using NDVI. Int. J. Rem. Sens. 22 (17), 3483–3501. http://dx.doi.org/10.1080/01431160010006430.

Lloyd-Huges, B., Saunders, M.A., 2002. A drought climatology for Europe. Int. J. Climatol. 22, 1571–1592.

Loukas, A., Vasiliades, L., 2004. Probabilistic analysis of drought spatiotemporal characteristics in Thessaly region, Greece. Nat. Hazards Earth Syst. Sci. 4 (5/6), 719–731.

Madadgar, S., Moradkhani, H., 2013. A Bayesian framework for probabilistic seasonal drought forecasting. J. Hydrometeorol. 14 (6), 1685–1705.

Mallya, G., Tripathi, S., Kirshner, S., Govindaraju, R., 2012. Probabilistic assessment of drought characteristics using a hidden Markov model. J. Hydrol. Eng. http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0000699.

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

McKee, T.B., Doesken, N.J., Kleist, J., 1995. Drought monitoring with multiple time scales. In: Proceedings of the Ninth Conference on Applied Climatology. American Meteorological Society, Dallas, TX, pp. 233–236.

Mishra, A.K., Singh, V.P., 2010. A review of drought concepts. J. Hydrol. 391 (1–2), 202–216. http://dx.doi.org/10.1016/j.jhydrol.2010.07.012.

Mishra, A.K., Desai, V.R., Singh, V.P., 2007. Drought forecasting using a hybrid stochastic and neural network model. J. Hydrol. Eng. 12 (6), 626–638. http://dx.doi.org/10.1061/(ASCE)1084-0699(2007) 12:6(626.

Mishra, A.K., Singh, V.P., Desai, V.R., 2009. Drought characterization: a probabilistic approach. Stochastic Environ. Res. Risk Assess. 23 (1), 41–55.

AghaKouchak, A., 2014. A baseline probabilistic drought forecasting framework using standardized soil moisture index: application to the 2012 United States drought. Hydrol. Earth Syst. Sci. 18 (7), 2485–2492.

Pan, M., Yuan, X., Wood, E.F., 2013. A probabilistic framework for assessing drought recovery. Geophys. Res. Lett. 40 (14), 3637–3642.

Pappenberger, F., Beven, K.J., 2006. Ignorance is bliss: or seven reasons not to use uncertainty analysis. Water Resour. Res. 42 (5), W05302. http://dx.doi.org/10.1029/2005WR004820.

Parathasarathy, B., Munot, A., Kothawale, D., 1994. Droughts over homogeneous regions of India: 1871–1990. Drought Network News 1994–2001, 67.

Rajeevan, M., 2006. High resolution daily gridded rainfall data for the Indian region: analysis of break and active monsoon spells. Curr. Sci. 91 (3), 296.

Richardson, S., Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). J. R. Stat. Soc.: Series B (Statistical Methodology) 59 (4), 731–792. http://dx.doi.org/10.1111/1467-9868.00095.

Rossi, G., Cancelliere, A., 2003. At-site and regional drought identification by REDIM model. In: Rossi, G., Cancelliere, A., Pereira, L.S., Oweis, T., Shatanawi, M., Zairi, A. (Eds.), Tools for Drought Mitigation in Mediterranean Regions. Water Science and Technology Library, Springer, Netherlands, pp. 37–54.

Russo, S., Dosio, A., Sterl, A., Barbosa, P., Vogt, J., 2013. Projection of occurrence of extreme dry-wet years and seasons in Europe with stationary and nonstationary standardized precipitation indices. J. Geophys. Res.: Atmos. 118 (14), 7628–7639. http://dx.doi.org/10.1002/jgrd.50571.

Ryu, J.H., Svoboda, M.D., Lenters, J.D., Tadesse, T., Knutson, C.L., 2010. Potential extents for ENSO-driven hydrologic drought forecasts in the United States. Climatic Change 101 (3–4), 575–597. http://dx.doi.org/10.1007/s10584-009-9705-0.

Shiau, J.-T., Feng, S., Nadarajah, S., 2007. Assessment of hydrological droughts for the Yellow River, China, using copulas. Hydrol. Process. 21 (16), 2157–2163. http://dx.doi.org/10.1002/hyp.6400.

Song, C., 2011. Report on the 2011 Symposium of Data-Driven Approaches to Droughts.

Sprague, L.A., 2005. Drought effects on water quality in the South Platte River Basin, Colorado. J. Am. Water Resour. Assoc. 41 (1), 11–24.

Steinemann, A., 2003. Drought indicators and triggers: a stochastic approach to evaluation. JAWRA J. Am. Water Resour. Assoc. 39 (5), 1217–1233.

Wiper, M., Insua, D.R., Ruggeri, F., 2001. Mixtures of gamma distributions with applications. J. Comput. Graphical Stat. 10 (3), 440–454. http://dx.doi.org/10.1198/106186001317115054.