DROUGHT CHARACTERIZATION USING PROBABILISTIC MODELS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Ganeshchandra Mallya

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2020

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF APPROVAL

Dr. Rao S. Govindaraju, Co-Chair

Lyles School of Civil Engineering, Purdue University

Dr. Shivam Tripathi, Co-Chair

Indian Institute of Technology Kanpur, India

Dr. Dennis Lyn

Lyles School of Civil Engineering, Purdue University

Dr. Bernard Engel

Department of Agricultural & Biological Engineering, Purdue University

Dr. Dev Niyogi

Department of Earth, Atmospheric, and Planetary Sciences, Purdue University

Approved by:

Dr. Dulcy Abraham

Head of the School Graduate Program

ACKNOWLEDGMENTS

I express my sincere gratitude to my advisor Dr. Rao S. Govindaraju, for his valuable guidance, patience, and continuous support throughout my Ph.D.

I would also like to thank Dr. Shivam Tripathi for being a great mentor throughout my graduate studies. Even though he was half-way around the world, he always found time to talk to me about my work and personal well being - and I will always be grateful to him.

I am honored to have Dr. Dennis A. Lyn, Dr. Dev Niyogi, and Dr. Bernard Engel as my committee members. I thank them for having have played a key role in my graduate education and continuously motivating me in my research by sharing their ideas and suggesting improvements to my work.

The staff at the Lyles School of Civil Engineering are the best in the world, and I would like to specially express my gratitude to Carmen Turner, Debra Burrow, and Jennifer Ricksy for being caring and helpful. I thank my wonderful colleagues in Hydraulics and Hydrology department for all their support, friendship, and sharing some fun moments.

Finally, and most importantly, I would like to thank my wife Sridevi, for her support, patience, and love. This journey wouldn't have been easy if it was not for her.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

ABSTRACT

Mallya, Ganeshchandra Ph.D., Purdue University, August 2020. Drought Characterization using Probabilistic Models. Major Professor: Rao S. Govindaraju.

Droughts are complex natural disasters caused due to deficit in water availability over a region. Water availability is strongly linked to precipitation in many parts of the world that rely on monsoonal rains. Recent studies indicate that the choice of precipitation datasets and drought indices could influence drought analysis. Therefore, drought characteristics for the Indian monsoon region were reassessed for the period 1901-2004 using two different datasets and standard precipitation index (SPI), standardized precipitation-evapotranspiration index (SPEI), Gaussian mixture model-based drought index (GMM-DI), and hidden Markov model-based drought index (HMM-DI). Drought trends and variability were analyzed for three epochs: 1901-1935, 1936-1970 and 1971-2004. Irrespective of the dataset and methodology used, the results indicate an increasing trend in drought severity and frequency during the recent decades (1971-2004). Droughts are becoming more regional and are showing a general shift to the agriculturally important coastal south-India, central Maharashtra, and IndoGangetic plains indicating food security challenges and socioeconomic vulnerability in the region.

Drought severities are 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. A probabilistic Gamma mixture model (Gamma-MM)-based drought index is proposed as an alternative to deterministic classification by SPI. The Bayesian framework of the

proposed model avoids over-specification and overfitting by choosing the optimum number of mixture components required to model the data − a problem that is often encountered in other probabilistic drought indices (e.g., HMM-DI). When sufficient number of components are used in Gamma-MM, it can provide a good approximation to any continuous distribution in the range $(0,\infty)$, thus addressing the problem of choosing an appropriate distribution for SPI analysis. The Gamma-MM propagates model uncertainties to drought classification. The method is tested on rainfall data over India. A comparison of the results with standard SPI shows significant differences, particularly when SPI assumptions on data distribution are violated.

Finding regions with similar drought characteristics is useful for policy-makers and water resources planners in the optimal allocation of resources, developing drought management plans, and taking timely actions to mitigate the negative impacts during droughts. Drought characteristics such as intensity, frequency, and duration, along with land-use and geographic information, were used as input features for clustering algorithms. Three methods, namely, (i) a Bayesian graph cuts algorithm that combines the Gaussian mixture model (GMM) and Markov random fields (MRF), (ii) k-means, and (iii) hierarchical agglomerative clustering algorithm were used to find homogeneous drought regions that are spatially contiguous and possess similar drought characteristics. The number of homogeneous clusters and their shape was found to be sensitive to the choice of the drought index, the time window of drought, period of analysis, dimensionality of input datasets, clustering method, and model parameters of clustering algorithms. Regionalization for different epochs provided useful insight into the space-time evolution of homogeneous drought regions over the study area. Strategies to combine the results from multiple clustering methods were presented. These results can help policy-makers and water resources planners in the optimal allocation of resources, developing drought management plans, and taking timely actions to mitigate the negative impacts during droughts.

1. INTRODUCTION

Droughts rank first among natural disasters or hazards in terms of the number of people directly affected and also in terms of accompanying socio-economic costs. According to Federal Emergency Management Agency (1995), the cost of droughts in the United States is estimated to be between $6 - 8 billion annually. Droughts occur when there is a precipitation (or other water sources) deficit compared to their long-term mean. Unlike other natural hazards such as floods or earthquakes, droughts are characterized by slow onset - often making it challenging to identify when a drought event begins. Also, droughts persist for longer durations, and their spatial extents are often large (Wang et al., 2011). The negative impacts of droughts over a region can be felt for an extended period, even after they have ended.

Droughts are classified into different categories based on their severity and are published regularly as bulletins in most countries. Water resources planners then use these drought classification bulletins to identify regions affected by drought, assess the amount of aid that these regions should receive based on the severity of drought, and also decide on drought mitigation strategies (Svoboda et al., 2002; Heim, 2002). Drought readiness schemes are also based on drought classification. While real-time drought monitoring is necessary for the timely allocation of required resources to mitigate the immediate negative impacts caused by droughts, decision-makers are interested in estimates of uncertainty in drought classification for developing informed remedial strategies. Although many sources of uncertainty exist, experts believe that allocation of resources and response capabilities of communities will benefit from the use of a drought index that provides estimates of model uncertainty through a probabilistic drought classification (Hayes et al., 2004; Song, 2011).

Global drought climatology has been assessed by different research groups to provide a baseline for future climate studies (Sheffield et al., 2012; Dai, 2013). The

primary inputs for any drought study are the measures of departures from long-term mean of variables governing water supply such as precipitation, soil moisture, runoff and streamflow. These inputs are available in several forms: i) point measurements at stations, ii) gridded datasets, often at different resolutions (e.g. 0.5°, 1°, etc.), obtained after interpolating point measurements from a single source or multiple sources that may include satellite based measurements, and iii) model outputs on a regular grid. These inputs are generally prone to measurement errors and model uncertainties that should be accounted for when used in further analyses. The next step is the choice of an indicator to measure the severity of droughts. The most popular drought indicators like Palmer drought severity index (PDSI) (Palmer, 1965) and standardized precipitation index (SPI) (McKee et al., 1993) do not account for model uncertainties and provide discrete drought classification. However, as noted in recent global drought climatology studies (Dai, 2013; Sheffield et al., 2012; Trenberth et al., 2014), the choice of input and other forcing datasets - in addition to the choice of model parameterizations used in drought indices - can lead to contrasting conclusions.

Droughts are also known to have a large spatial footprint. Thus, in addition to considering the time evolution of droughts, it is also important to consider their spatial characteristics. A multi-site analysis can characterize droughts over a region. A multi-site analysis implies that relevant input data (or drought characteristics) from neighboring stations or grids located within a region are pooled. The choice of neighboring stations (or grids) often tends to be subjective. Identifying drought-prone regions and regions with similar drought characteristics in an objective manner is important for drought management purposes. If a region is prone to frequent droughts, sufficient resources need to be allocated, and appropriate remedial measures need to be implemented by drought managers to improve the region's resilience to droughts. However, if such measures are not implemented in a timely fashion, it may lead to costly environmental degradation and eventually result in desertification. Thus, it is important to identify drought hotspots from the perspective of developing drought readiness schemes. The use of homogeneous regions in terms of overall climatol-

ogy (Karl and Koss, 1984; Bharath and Srinivas, 2015), or monsoon precipitation (Parthasarathy et al., 1993) are common in climate research. Several studies (Stahl and Demuth, 1999; Trnka et al., 2009; Vicente-Serrano, 2006a) have also focused on identifying homogeneous drought regions, i.e. regions with similar drought climatology. Dracup et al. (1980), suggested that stations within homogeneous drought regions should not only be similar in terms of climate but also be geographically contiguous and have similar geomorphology. Homogeneous drought regions provide an objective way to pool hydro-meteorological data for performing regional and multi-year drought analysis. The inter-relationships between homogeneous drought regions may be used to develop early drought warning systems.

The study objectives are as follows:

1. To perform retrospective drought analysis for evaluating trends and variability of drought events over the Indian monsoon region (IMR).

2. To develop a probabilistic drought index by accounting model and parameter uncertainties.

3. To develop a Bayesian framework that uses concepts of Gaussian mixture model (GMM) and Markov random fields (MRFs) for finding homogeneous drought regions that are spatially contiguous and have similar drought characteristics.

The thesis is structured as follows: in Chapter 2, a retrospective drought analysis over IMR using multiple data sources and methods will be presented. In Chapter 3, a probabilistic drought index that engages model uncertainty by providing probabilistic drought classes will be introduced. In Chapter 4, several data-driven unsupervised clustering methods will be used to identify regions with similar drought characteristics. Finally, in Chapter 5, the summary and conclusions of this study will be presented.

2. TRENDS AND VARIABILITY OF DROUGHTS OVER INDIAN MONSOON REGION

This article has been previously published in Weather and Climate Extremes June 2016.

2.1 Introduction

Droughts in the monsoon dominated regions have gained greater importance in the recent past, as monsoons not only define the unique features of the climate, but also affect the socioeconomic well-being of more than two-thirds of global population (Niranjan Kumar et al., 2013; Rajeevan et al., 2008). Recent changes in Indian monsoon precipitation have received wide attention (Kripalani et al., 2003; Mishra et al., 2012; Rupa Kumar et al., 2006) with some plausible uncertainty on whether trends associated with summer monsoon precipitation are related to global warming or regional changes (Chung and Ramanathan, 2006; Kishtawal et al., 2010; Niyogi et al., 2010). A number of studies (Kumar et al., 1992; Rajeevan et al., 2008; Stephenson, 2001) have indicated that the mean precipitation during the monsoon season may be unaltered over the Indian Monsoon Region (IMR), but the extreme precipitation events have shown statistically significant increasing trends in last five decades resulting in modification of drought characteristics over IMR (Goswami et al., 2006; Mishra et al., 2012). Trends associated with the Indian summer monsoon rainfall (ISMR) have also shown a great regional variability where some parts of India have seen an increase in precipitation while others show a reduction in precipitation during the monsoon season (Guhathakurta and Rajeevan, 2008; Niyogi et al., 2010; Roxy et al., 2015). Significant interannual, decadal and long term trends have been observed in the monsoon drought time series over IMR influenced by El Nino Southern Oscillation (ENSO) and global warming (Niranjan Kumar et al., 2013).

Recently, contrasting conclusions were drawn about global drought climatology by two synthesis studies (Sheffield et al., 2012; Dai, 2013). While Sheffield et al. (2012) showed that there was little change in drought climatology in recent years, the study by Dai (2013) concluded that droughts were intensifying as a result of a warming climate. Subsequently, Trenberth et al. (2014) summarized that the choice of precipitation dataset and other forcing datasets could influence drought analysis in addition to the choice of model parameterizations being used in deriving the drought indices [e.g. potential evapotranspiration calculations while estimating Palmer Drought Severity Index (PDSI) as reported in Sheffield et al. (2012)]. These studies highlight the need for using multiple drought indices and datasets for drought climatology, and form the basis for reassessing droughts over IMR.

Evaluation of trends and variability associated with retrospective drought events provides a basis to understand regional patterns of severity, duration, and areal extent of droughts. It also enables an understanding of the nature of possible future droughts and potential vulnerabilities. Building off the findings of drought assessments over IMR in recent years and the recommendations cited in Trenberth et al. (2014), the aims of this work are (i) to study retrospectively, the droughts and associated trends over IMR using different precipitation datasets and drought indices, and (ii) to identify regions in IMR that are vulnerable to droughts.

2.2 Data and Methods

Gridded daily precipitation data from the India Meteorological Department (IMD) (Rajeevan, 2006) available for the period $1901 - 2004$ at 1° spatial resolution (Figure 2.1) were used. The daily precipitation data obtained from IMD were then aggregated over monthly time scale. The second dataset used in this study was monthly precipitation data from University of Delaware (UD) available for the period of $1900 - 2004$ (UDel_AirT_Precip data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their website at http://www.esrl.noaa.gov/psd/) at 0.5° spatial

resolution (Figure 2.1). The precipitation data from high-mountainous regions in northern and northeastern parts of the country were not used in the study as the number of rain gauges in these regions are sparse.

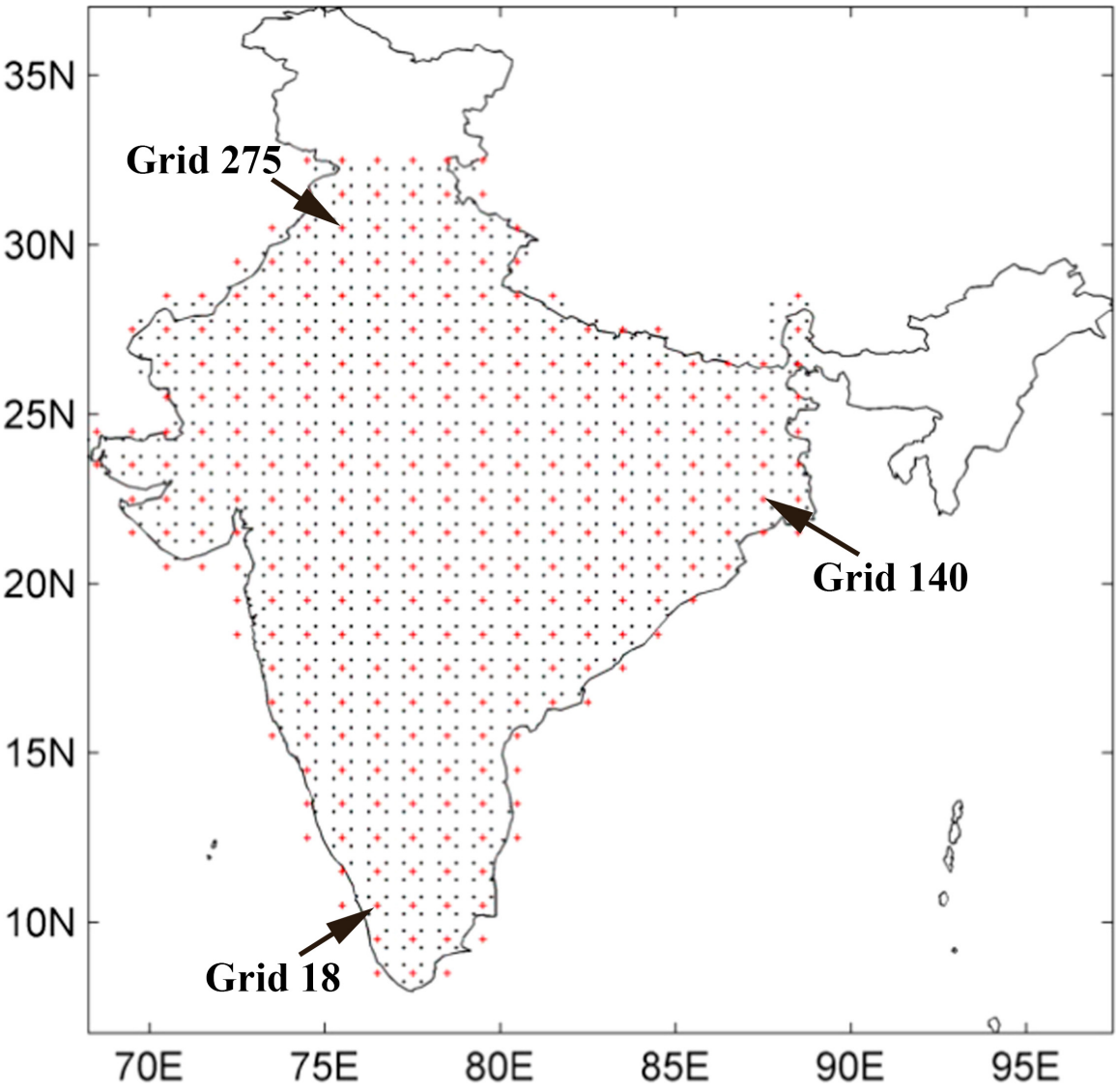

Figure 2.1. Study domain showing 1° grid cell locations for IMD precipitation dataset as cross-hairs, and 0.5° grid cell locations for UD precipitation dataset as dots.

Despite the differences in the spatial resolution, the precipitation datasets show similar patterns in the spatial distribution and variance of precipitation over the study

region. Figures 2.2a-b show the distribution of mean monthly precipitation over the study region, and Figures 2.2c-d compare the standard deviation in monthly mean precipitation between the two datasets. While the overall patterns are similar, the effects of resolution on the magnitudes are evident. For instance, the UD dataset provides more detail in the spatial distribution of precipitation statistics.

A comparison of monthly mean precipitation time series (Figure 2.3) between the two datasets shows that while the overall monthly time series pattern are similar, the precipitation magnitude for IMD grids are lower compared to UD grids during the months June to September, and relatively identical for the remaining months.

Standardized precipitation index (SPI) (McKee et al., 1993), standardized precipitation evapotranspiration index (SPEI) (Vicente-Serrano et al., 2010; Niranjan Kumar et al., 2013), Gaussian mixture model-based drought index (GMM-DI) (Mallya, 2011), and Hidden Markov model-based drought index (HMM-DI) (Mallya, 2011; Mallya et al., 2013) were calculated for drought characterization at multiple time scales ending in September (i.e. for 1-month, 4-month, and 12-month moving time-window) and December (i.e. for 7-month moving time-window). The results for 12-month moving time window accounts for precipitation events occurring over both the monsoon and the non-monsoon months and 7-month time-window ending in December accounts for summer monsoon (JJAS) and winter monsoon (OND) months over the study area and are discussed here in detail. These indices differ in their mathematical formulation and the drought classification technique. While SPI relies on fixed thresholds for drought classification, GMM-DI and HMM-DI employ a probabilistic data-driven approach. SPEI uses temperature (UDel_AirT_Precip, http://www.esrl.noaa.gov/psd/) for calculating evapotranspiration, thus accounting for any temperature rise in the study area during recent decades. The mathematical formulations of the drought indices are summarized at the end of this section (refer to subsections 2.2.1 to 2.2.4).

The maximum drought time window considered in this study was 12-months. For longer time windows, non-overlapping data available at any grid point over the study

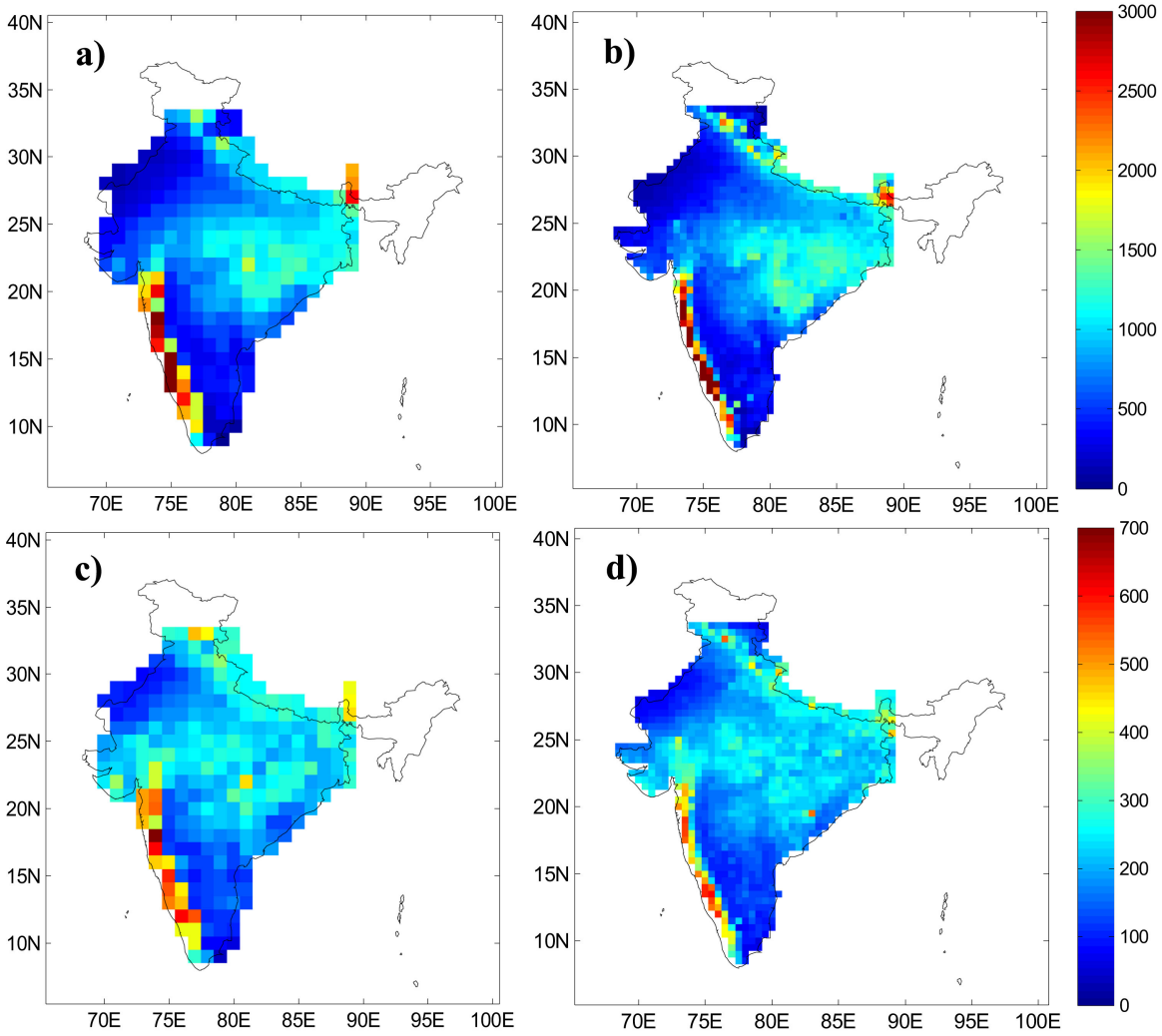

Figure 2.2. Comparison of IMD and UD monthly precipitation statistics over each grid in the study region. Mean of monthly precipitation (in mm) over (a) IMD grids, and (b) UD grids. Standard deviation of monthly precipitation (in mm) over (c) IMD grids, and (d) UD grids.

region would become limited. For analyzing multi-year droughts, data can be pooled from stations with similar drought characteristics. Most statistical models used to compute drought indices assume the data to be independent. Autocorrelation in drought time series can be computed for different time windows of interest to validate the assumption of independence. Figure 2.4 shows the auto-correlation in the SPI

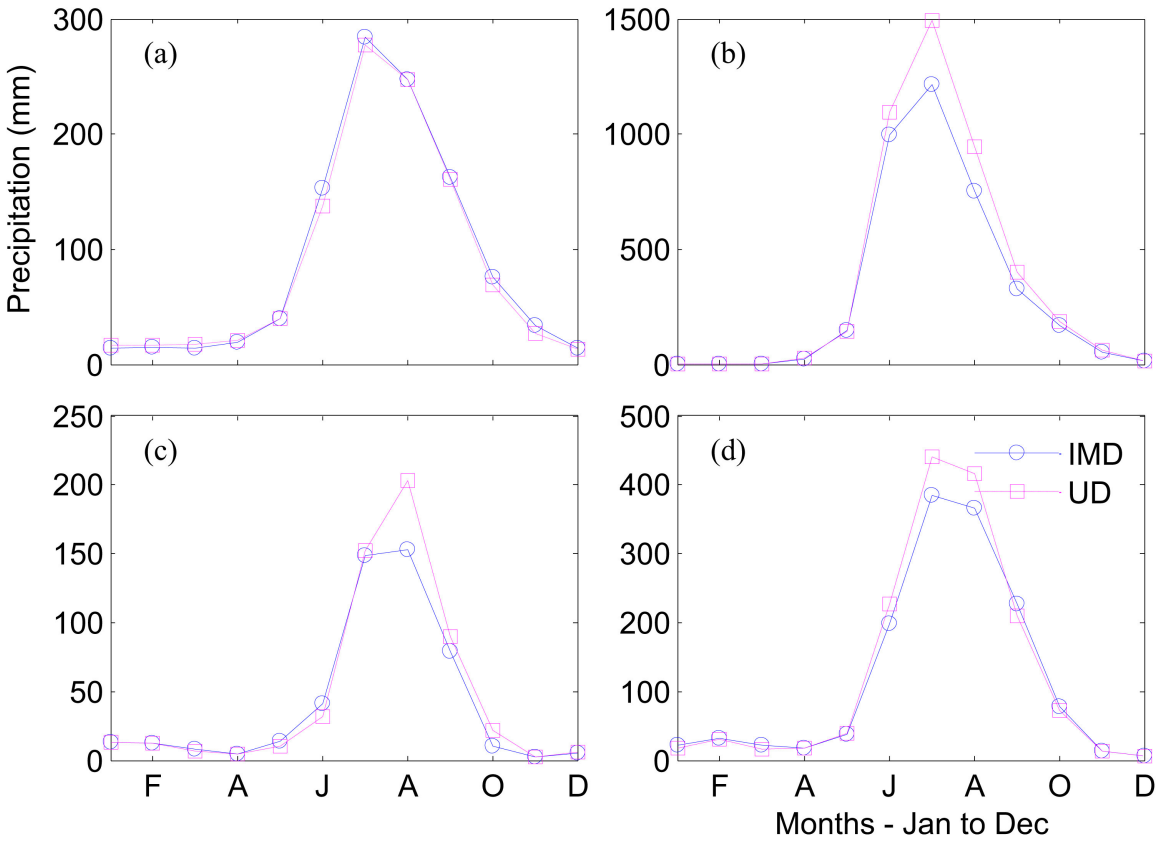

Figure 2.3. Comparison of monthly mean precipitation between IMD and UD datasets for all months of a year (January to December), (a) averaged over all grids in the study region, (b) IMD grid 18 and UD grid 208 over Western Ghats, (c) IMD grid 275 and UD grid 430 over Punjab, and (d) IMD grid 140 and UD grid 1005 over West Bengal. Location of IMD grids are shown in Figure 2.1. UD grids are located adjacent to IMD grids.

time series at different grids over the study region. Autocorrelation values close to zero indicate that the data are independent, while those closer to one indicate dependence. For a 12-month time window ending in May, Figures 2.4a-d show the auto-correlation at lags 1 to 4 months, respectively. Except for few grids over Jammu, Kashmir, northeastern states, and the Western Ghats along the west coast of India, the values of autocorrelation are close to zero starting at lag-1 and show a further decrease for higher lags. Similarly, observation in auto-correlation values may be made for 4-month

window ending in September (Figures 2.4e-h). The magnitude of auto-correlation was smaller across all lags in the 4-month time window, compared to the 12-month time window over grids in the study region. When the independence assumption is violated, a drought index like HMM-DI can be used to model the underlying data's dependence structure through its hidden states and revealed through the transition matrix.

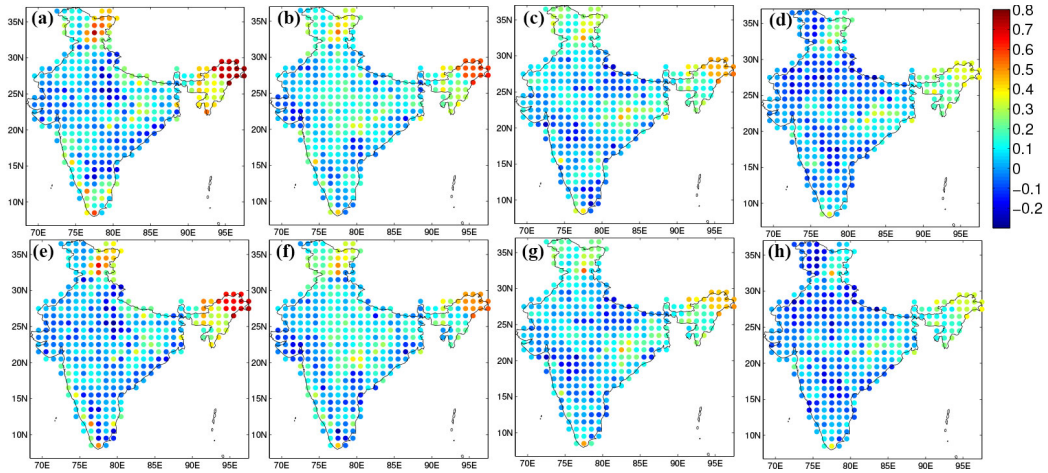

Figure 2.4. Autocorrelation in SPI time series. Top panel corresponds to a 12-month window ending in May (water year) for the study period 1901-2004 (P4) for (a) 1-month (b) 2-months (c) 3-months and (d) 4-months lag. Similarly, the bottom panel corresponds to a 4-month window ending in September.

The drought index values obtained were analyzed further to extract drought characteristics such as severity, duration, areal extent, and frequency. The drought impact index was then computed for each year, by normalizing the product of mean severity and the areal extent of drought.

The study period was divided into three segments (1901- 1935, 1936 -1970, and 1971 -2004) to understand the trends and variability associated with retrospective droughts. This was done because droughts have a multiyear influence, and the three periods chosen, approximately correspond to periods where IMR experienced significant droughts (e.g. 1918, 1965, 1972, 1979, 1987, and 2002; see Figure 2.5). Dividing

the entire 104 years (1901-2004) of data into three periods (35, 35, and 34 years) was expected to provide a sufficient length of time series to estimate trends and other statistical values.

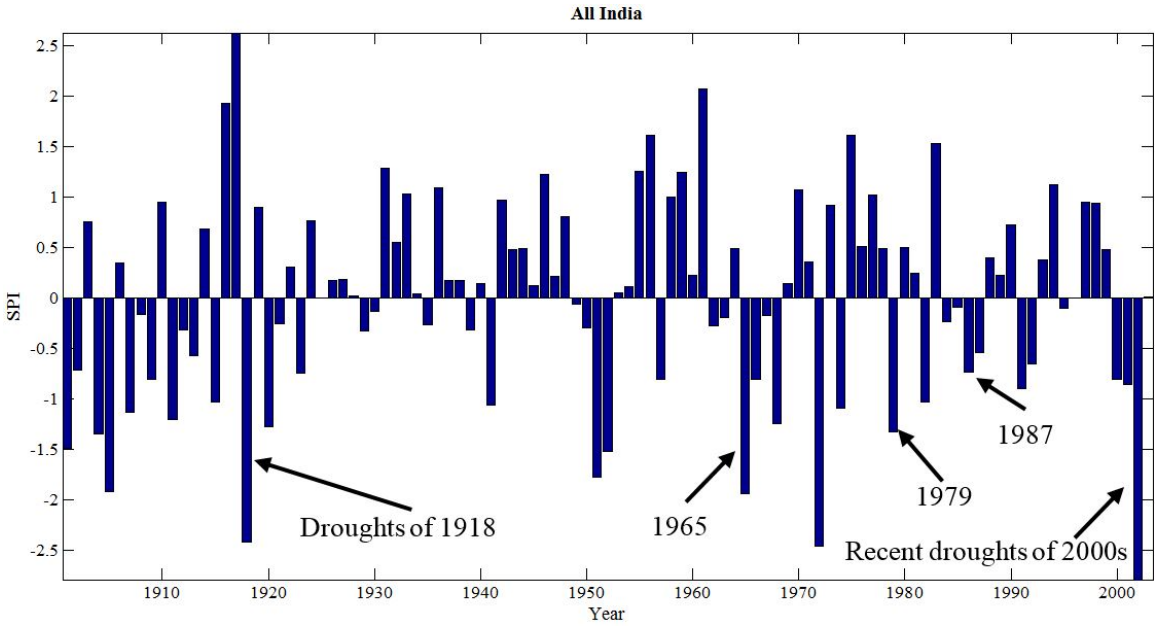

Figure 2.5. SPI corresponding to water year (June to May) over the Indian monsoon region

A modified Mann-Kendall trend test that accounts for autocorrelation in time-series (Kulkarni and von Storch, 1995; Hamed and Ramachandra Rao, 1998) was used to detect trends in the annual SPI, SPEI, GMM-DI, and HMM-DI values. Trends were estimated on the annual time series for the entire period and for each of the sub-periods (i.e. 1901-1935, 1936-1970, and 1971-2004) using a 5% significance test. The effect of spatial correlations in the data (Burn and Elnur, 2002; Yue and Wang, 2002) on the trend results was accounted for by using false discovery rate (FDR) (Benjamini and Hochberg, 1995; Ventura et al., 2004).

The mathematical formulations of the drought indices are briefly described below:

2.2.1 Standardized precipitation index (SPI)

The SPI measures the deficit in observed precipitation (McKee et al., 1993) and has been used widely to identify meteorological, agricultural, and hydrological droughts (Mishra and Singh, 2010; Mo, 2008). Precipitation time-series for each grid cell over IMR, at any desired time-scale, was first used to fit a probability distribution function, and then normalized using a standard inverse Gaussian function to obtain SPI values. Drought severity was identified using the SPI ranges as described by Charusombat and Niyogi (2011). A drought event was classified as moderate if SPI was between -1.0 to -1.49, severe if SPI was between -1.5 to -2.0, and extreme if SPI was less than -2.0.

2.2.2 Standardized precipitation evapotranspiration index (SPEI)

SPEI (Vicente-Serrano et al., 2010) first requires the computation of potential evapotranspiration (PET). Thornthwaite's equation (Thornthwaite, 1948) was used for computing PET, but other popular approaches may also be used (Penman, 1948; Priestley and Taylor, 1972; Allen et al., 1998). After subtracting the PET from precipitation, SPEI may be computed using similar approach as SPI (McKee et al., 1993).

2.2.3 Gaussian mixture model-based drought index (GMM-DI)

A GMM is a probabilistic model where the parametric density function is represented as a weighted sum of the Gaussian component densities (Reynolds and Rose, 1995). GMMs have been successfully used in recent studies involving estimation of weather parameters using remotely sensed radar data (Li and Zhang, 2011), and hydrologic forecasting studies (Liang et al., 2011). In this study, each individual Gaussian component is assumed to represent the underlying distribution of the hidden drought (or wet) classes with a mean μ, and a covariance matrix Σ. Such a

GMM model closely represents a hidden Markov model (HMM) with equal transition probabilities amongst all hidden states. The mathematical formulation of the GMM used in this study is described below.

Let the precipitation at time t be denoted by x_t, $t = 1 \ldots N$, $\{x_t \in R$ and $X = [x_1, \ldots, x_N]^T = x_{1:N}\}$. If the total number of components of the GMM, M, are known *a priori*, then the weighted sum of M component GMM is given by the equation,

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i) \tag{2.1}$$

where w_i are the mixture weights, and $g(x|\mu_i, \Sigma_i)$ are the component Gaussian densities of the form,

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{\frac{1}{2}(x-\mu_i)' \Sigma_i^{-1} (x-\mu_i)\right\} \tag{2.2}$$

with mean μ_i and covariance matrix Σ_i. In this study, since only precipitation data are used, the number of dimensions, $D = 1$. Further, the mixture weights satisfy the constraint $\sum_{i=1}^{M} w_i = 1$. The parameter set can be represented by the notation shown below.

$$\lambda = \{w_i, \mu_i, \Sigma_i\}, \; i = 1, \ldots, M \tag{2.3}$$

Expectation-maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997) was used to estimate the parameters of the GMM using a maximum likelihood approach. The *a posteriori* probability for component i was given by

$$p(i|x_t, \lambda) = \frac{w_i p(x_t|\mu_i, \Sigma_i)}{\sum_{k=1}^{M} w_k p(x_t|\mu_k, \Sigma_k)} \tag{2.4}$$

To compare the results of GMM-DI with SPI, the number of mixture components, M, was set to 7 (3 drought states + 1 normal state + 3 wet states).

2.2.4 Hidden Markov model-based drought index (HMM-DI)

The mathematical formulation of the Hidden Markov model-based drought index (HMM-DI) is described in detail in Mallya (2011). Again, let precipitation 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 a HMM, the precipitation x_t is assumed to depend only on the state variable $z_t,\ \left\{Z = [z_1, \ldots, z_N]^T\right\}$ that denotes a drought or wet state, is hidden (not observed), and follows the first order Markov property. The state variable z_t is a K-dimensional binary random variable. If the number of states, K, are known *a priori*, the standard HMM can be parameterized using the following three distributions

i The conditional distribution of precipitation given the drought state, $p(x_t|z_t)$, referred to as the emission distribution;

ii The conditional distribution of the present drought state given the previous state i.e. $p(z_t|z_{t-1})$. Because z_t is a K dimensional binary variable, the conditional distribution is given by a $K \times K$ transition matrix A whose element $A_{jk} = p(z_{tk} = 1|z_{t-1,j} = 1)$;

iii The marginal distribution of the drought state at the first time step, $p(z_1)$, is given by a K dimensional vector Π whose element $\pi_k = p(z_{1k} = 1)$.

The precipitation data at the desired time-scale was transformed to represent percentage deviation from their long term mean. The HMM model was then applied to this transformed data. The probability density function of the emission distribution was selected to be a Gaussian distribution of the form

$$p(x_t|z_t) = \prod_{k=1}^{K} N\left(x_t|\mu_k, \sigma_k^2\right)^{z_{tk}} \tag{2.5}$$

where μ_k and σ_k^2 are the mean and the variance of a Gaussian distribution, respectively. The μ_k's and σ_k's were considered to be free parameters and were estimated

along with other parameters of HMM. Since the results of HMM-DI are to be compared with SPI and GMM-DI, the number of states (components in the Gaussian mixture) K was set to 7 (3 drought states + 1 normal state + 3 wet states).

The underlying emission distributions are not known before hand and were assumed to be Gaussian. This was done for mathematical convenience, and also because many processes combine to create droughts, and one may expect that their combined influence expressed through deviations from the mean could be Gaussian. Additionally, if there is no temporal dependence, the HMM automatically collapses to a Gaussian mixture for which the theories are well developed.

Both GMM-DI and HMM-DI provide probabilities of belonging to each drought class. To obtain drought intensity values, the intensity factors (e.g. assumed intensity factors for Extreme drought = -3.0, Severe drought = -2.0, Moderate drought = -1.0 and so on) were multiplied with corresponding probability measures.

2.3 Results and discussion

2.3.1 Drought characterization

The drought indices were able to capture (Figure 2.6, Figure 2.7a and $1^\circ \times 1^\circ$ IMD dataset) the major documented drought events over IMR (De et al., 2005). For the study period, the six most notable moderate-droughts occurred in 1905, 1946, 1965, 1974, 1979, and 1984. In the figures, moderate drought refers to SPI values between -1.0 and -1.49 (Charusombat and Niyogi, 2011). Three of the most severe historic droughts occurred during the recent period of 1971-2004. The drought characteristics showed an increasing trend during the same period. Modified Mann-Kendall trend test was performed to test the statistical significance of the trends in these average drought statistics. For example, SPI analyses (Figure 2.6a) showed an increasing trend in the mean severity of moderate droughts (-0.04/decade, trend is towards negative SPI values, p-value > 0.05) during the period 1971-2004. During the same period, the areal extent and drought impact index of moderate droughts also showed

increasing trends. Similar trends were observed for SPEI, GMM-DI and HMM-DI analyses (Figure 2.7a, Figure 2.6b-2.6c). These trends are consistent with the precipitation trends documented in other studies (Guhathakurta and Rajeevan, 2008; Kripalani et al., 2003; Rupa Kumar et al., 2006).

The trends were reanalyzed in the 0.5° resolution UD precipitation data, thus providing means to compute and validate trends in drought characteristics at a relatively finer spatial resolution over IMR. As in the case of IMD dataset, SPI, SPEI, GMM-DI and HMM-DI were computed. Drought characteristics such as mean severity, areal extent, and drought impact index were computed for each drought index. Again the drought indices were able to capture (Figure 2.8, Figure 2.7b) the major drought events documented over IMR (De et al., 2005) during the period of 1901-2004 and agree well with IMD dataset results (Figure 2.6, Figure 2.7a). There are broad similarities and also specific differences in the characteristics revealed by the choice of index and data. For example, the SPI and SPEI yields a relatively smaller drought impact index as compared to GMM-DI and HMM-DI.

Figure 2.6. Drought characteristics over IMR computed for IMD dataset using (a) SPI, (b) GMM-DI, and (c) HMM-DI for 12-month time window ending in September. In each figure the top-panel shows time-series plot of moderate drought severity averaged over all grids. Middle-panel shows the bar-plot of areal extent of moderate droughts represented as percentage of total area in the IMR. Bottom-panel shows the bar-plot of drought impact index for moderate droughts. Solid line represents the median value and dotted line represents slope during the sub-periods 1902-35, 1936-70 and 1971-2004 respectively.

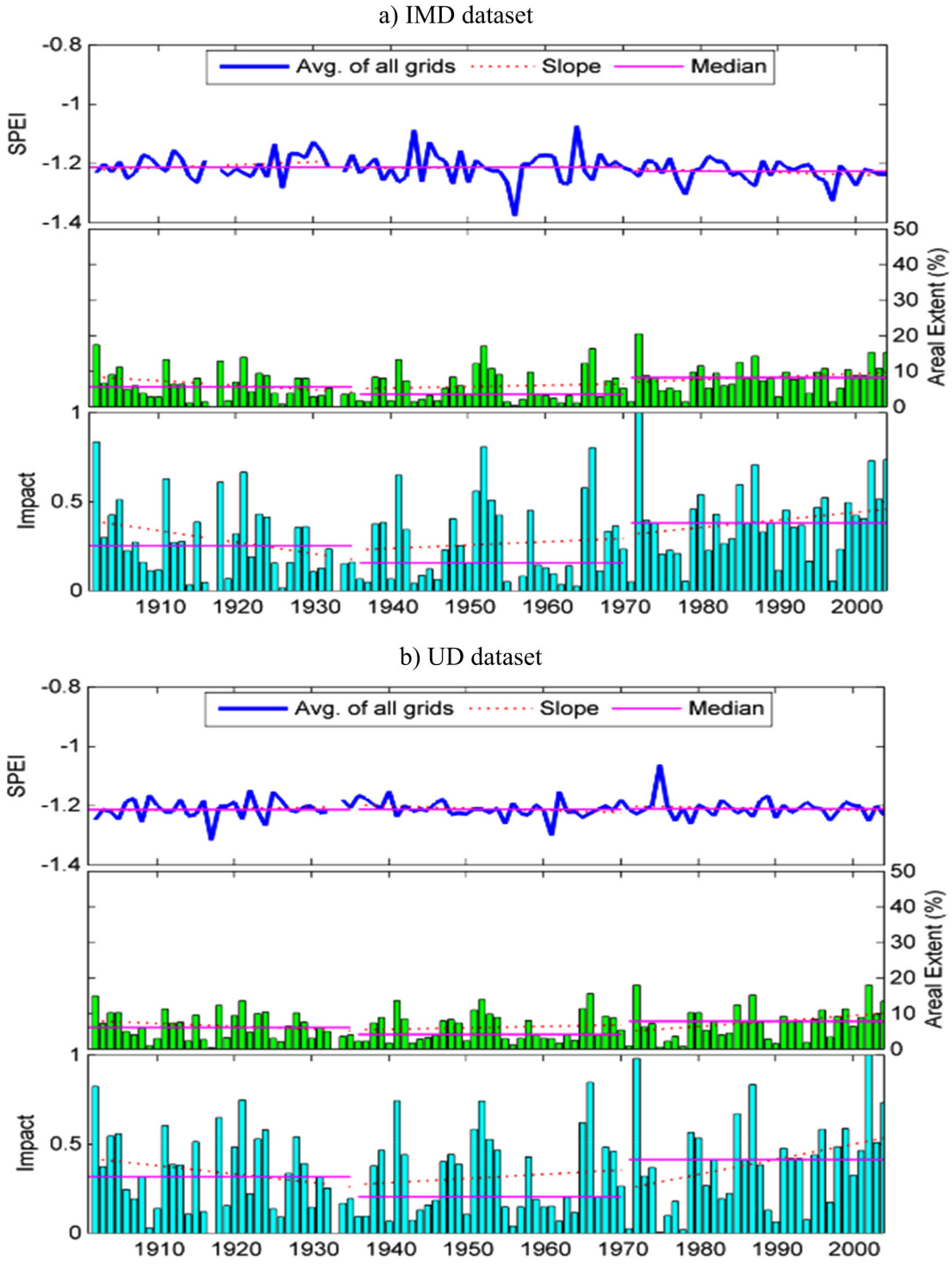

Figure 2.7. Same as Figure 2.6, but for SPEI using (a) IMD and (b) UD datasets, respectively.

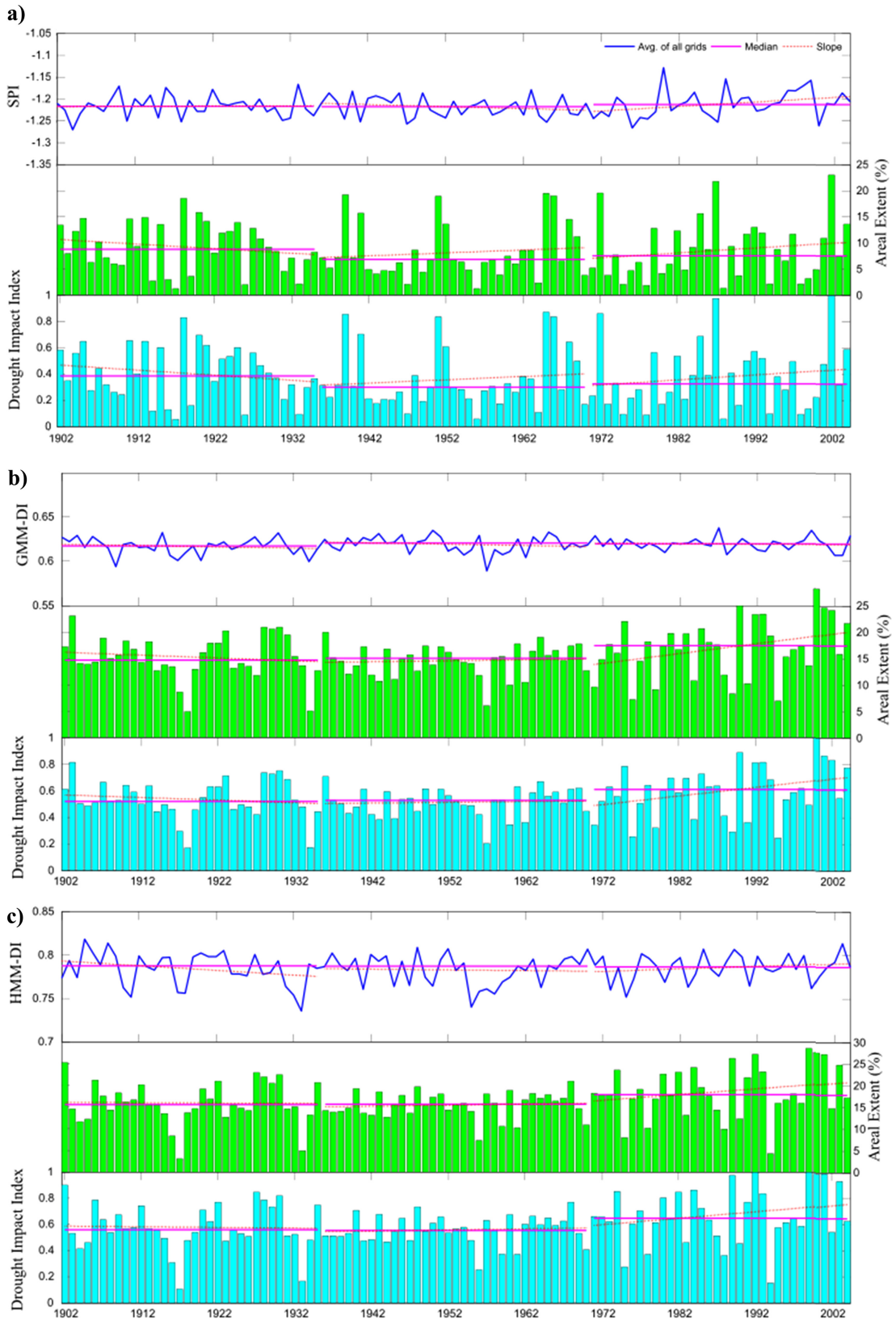

Figure 2.8. Same as Figure 2.6, but using 0.5° UD dataset

2.3.2 Spatial and temporal variability in drought characteristics

To study the spatiotemporal variability in droughts, average drought characteristics based on SPI, SPEI, GMM-DI, and HMM-DI values were obtained for each epoch over all grids in IMR by computing the mean number, severity and duration of droughts (e.g. 1901-1935; 1936-1970 and 1971-2004). For the IMD dataset and 12-month time window, during the period 1901-35 there were many widespread droughts (Figures 2.9-2.11) mainly in the northern, central and the Deccan Plateau regions of IMR. While more drought events were observed in the Deccan region (Figure 2.9), the drought duration (Figure 2.11) and intensity (Figure 2.10) were higher in northern and central regions of IMR. During the epoch of 1936-1970, droughts were more active in the western region and parts of Deccan Plateau of IMR. Compared to 1901-35, droughts were less frequent during this epoch (1936-1970). During 1971-2004, the number of drought events and their duration increased in the central and eastern Indo-Gangetic plain (IGP) (20°N-28°N), and southern parts of IMR. High drought intensities were recorded in central and eastern IGP, south-India, and parts of western-India (that include states of Maharashtra, Gujarat, and Rajasthan). Drought patterns were mostly similar for all four drought indices in each epoch - while GMM-DI showed more widespread droughts; SPI, SPEI, and HMM-DI were better able to distinguish the drought hotspots.

Results for the UD dataset were similar to those obtained for IMD dataset. There were many widespread droughts in the western and central parts of IMR during 1901-1935 (Figures 2.12-2.14). During 1936-1970, except for some parts of western, central and southern India, most of the IMR was wet and droughts were infrequent. As in case of IMD dataset (Figures 2.9-2.11), the number of droughts and duration of droughts increased in the central and eastern IGP (20°N-28°N), and southern parts of India during 1971-2004. The drought intensities were higher in interior parts of south-India, western parts of India (Maharashtra, Gujarat, and Rajasthan) and IGP. The drought indices - SPI, SPEI, GMM-DI and HMM-DI - were able to consistently

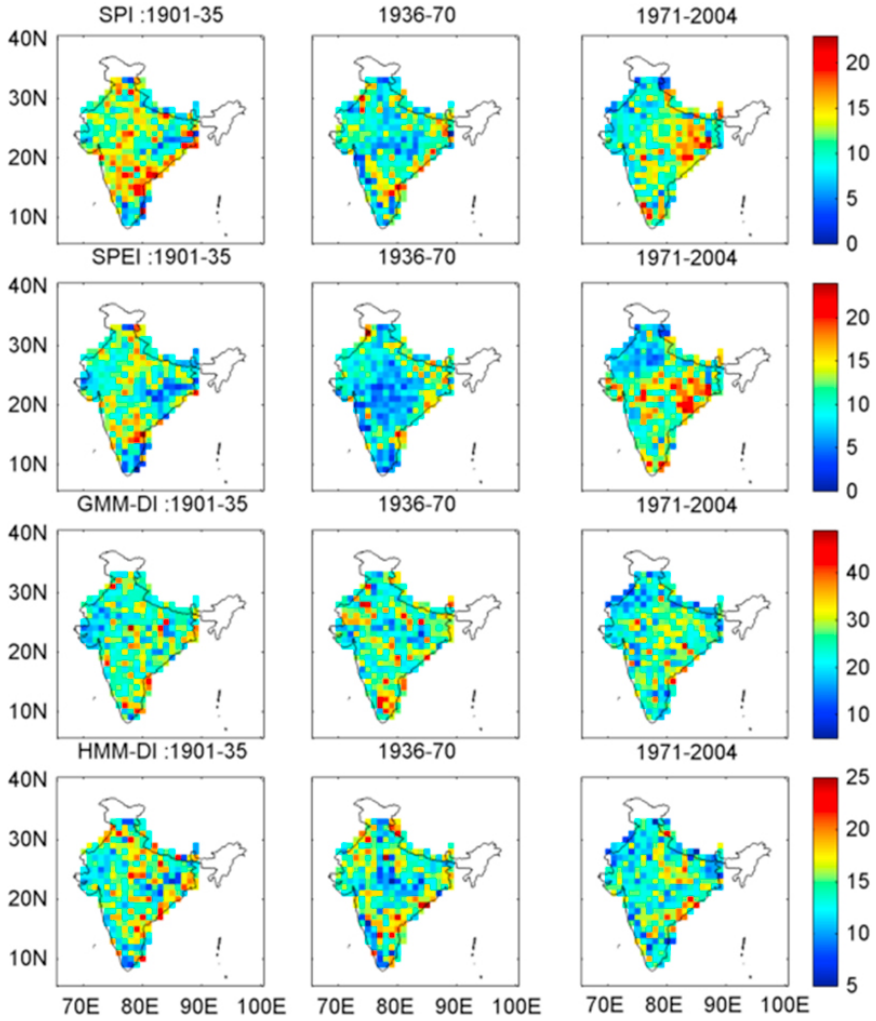

Figure 2.9. Epochal variation in number of drought events over IMR using IMD dataset. In each sub-plot top panel represents SPI, followed by SPEI, GMM-DI, and HMM-DI.

capture the space and time evolution of drought characteristics over the IMR during the entire study period. A notable west to east migration in the drought severity and extent over the last century is observed from Figures 2.9 to 2.14.

Similar comparison of epochal drought characteristics over IMR for 7-month time window using IMD (Figure 2.15) and UD (Figure 2.16) datasets showed that during 1901-1935 droughts were more intense and frequent in parts of Deccan Plateau,

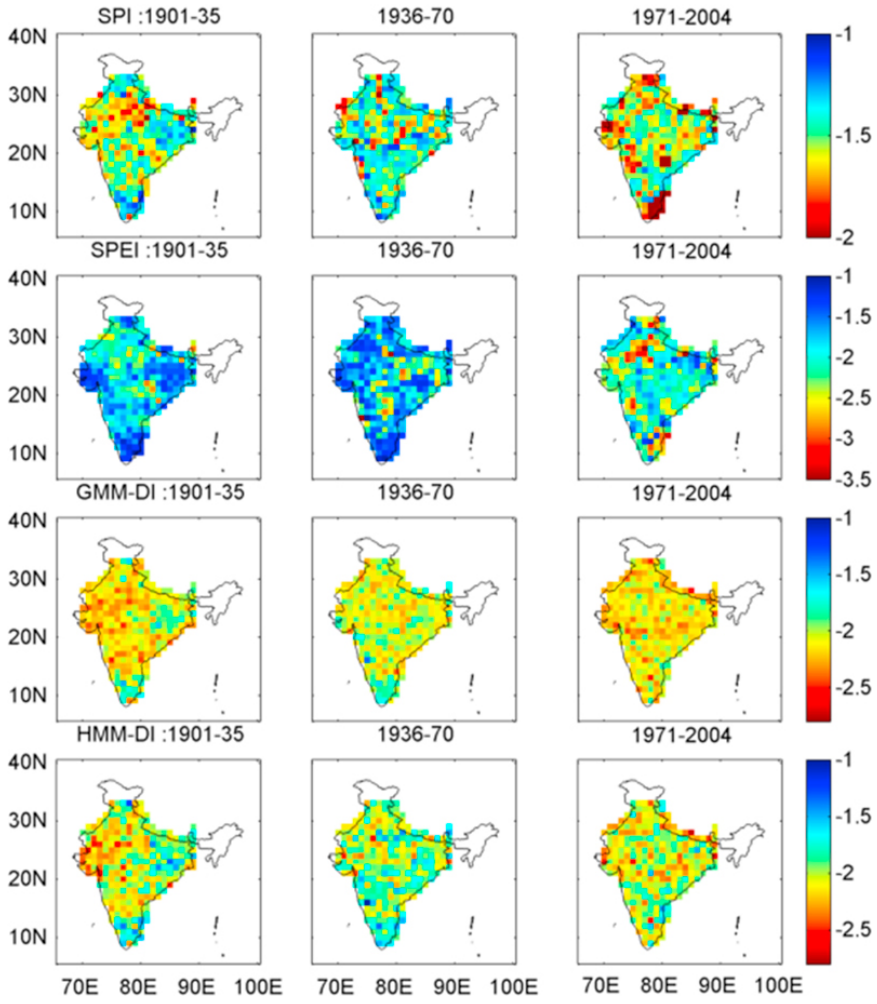

Figure 2.10. Same as Figure 2.9, but for drought intensity.

western and northern parts of India. Droughts were comparatively less frequent during the epoch 1936-1970 according to SPI and SPEI, however GMM-DI and HMM-DI analysis shows that droughts continue to be intense and frequent in western India and parts of Deccan Plateau. During 1971-2004 central-India, eastern IGP, and parts of south-India emerge as drought hotspots - along with high intensity but short-term droughts in western-India.

A decadal comparison of 12-month time window drought characteristics over IMR using IMD dataset (Figure 2.17-2.19) shows higher level of drought activity

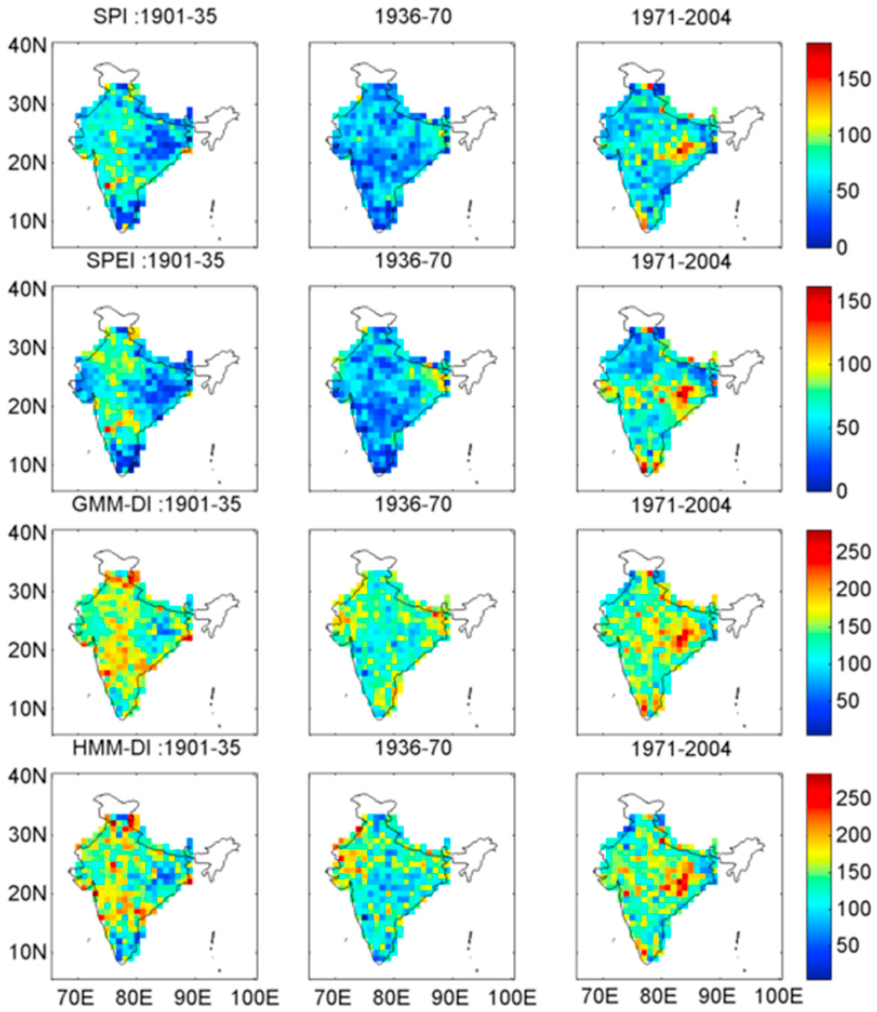

Figure 2.11. Same as Figure 2.9, but for drought duration (in months).

in northern-India, western-India and Deccan Plateau during the 1901-10, 1911-20, with more intensification during 1921-30. The subsequent two decades (1931-40 and 1941-50) were amongst the wettest in the past century. Droughts started to emerge in the eastern-IGP during late 1951-60 and intensified in IGP and parts of western-India during 1961-70. During 1971-80 droughts continued to persist over eastern IGP, and in the following decade (1981-90) additional hotspots emerged in south-India and parts of western-India. During 1991-2000 and onwards, eastern-IGP and parts of central-India continue to be drought hotspots. Similar patterns in drought charac-

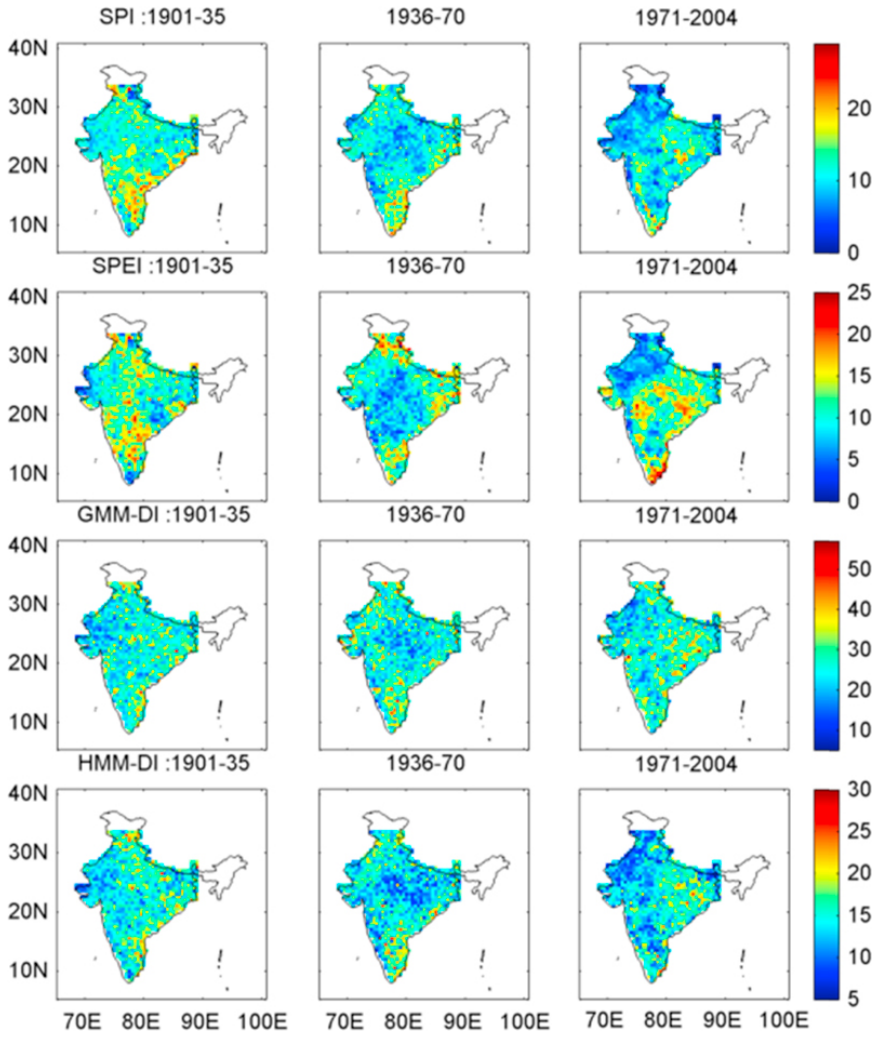

Figure 2.12. Epochal variation in number of drought events over IMR using UD dataset. In each sub-plot top panel represents SPI, followed by SPEI, GMM-DI, and HMM-DI.

teristics were observed in our analysis when using UD dataset, and for different time windows.

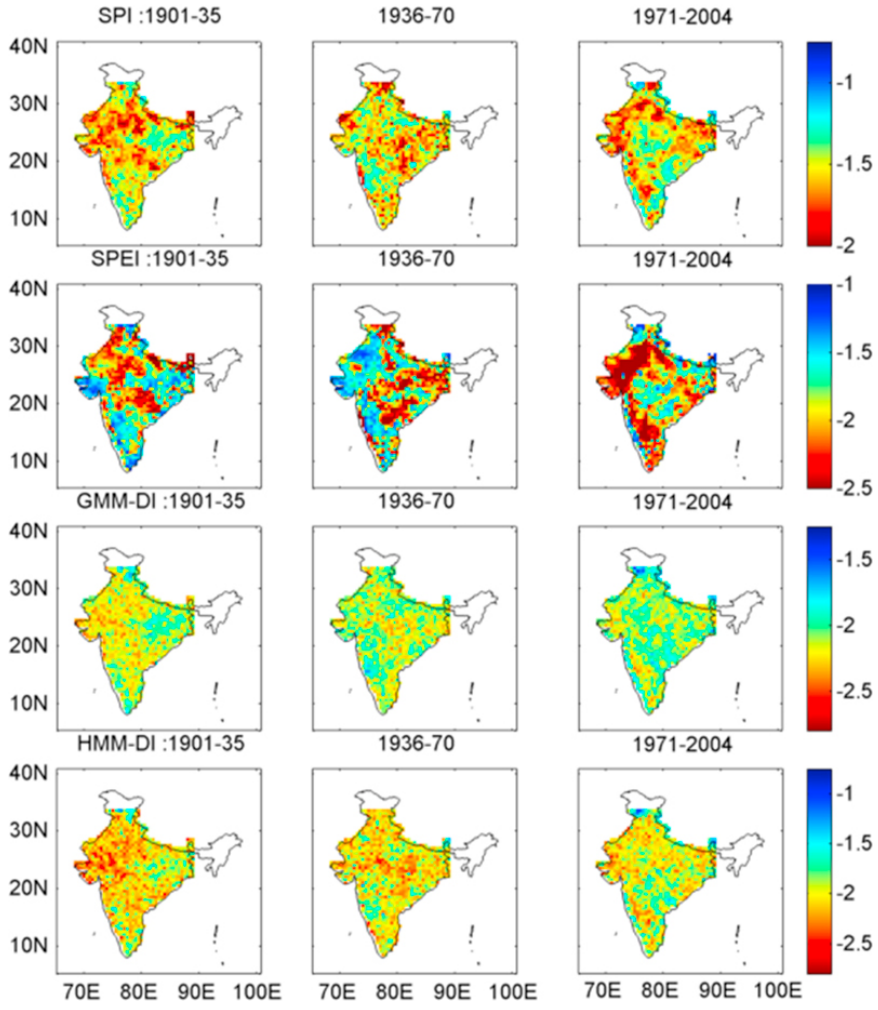

Figure 2.13. Same as Figure 2.12, but for drought intensity.

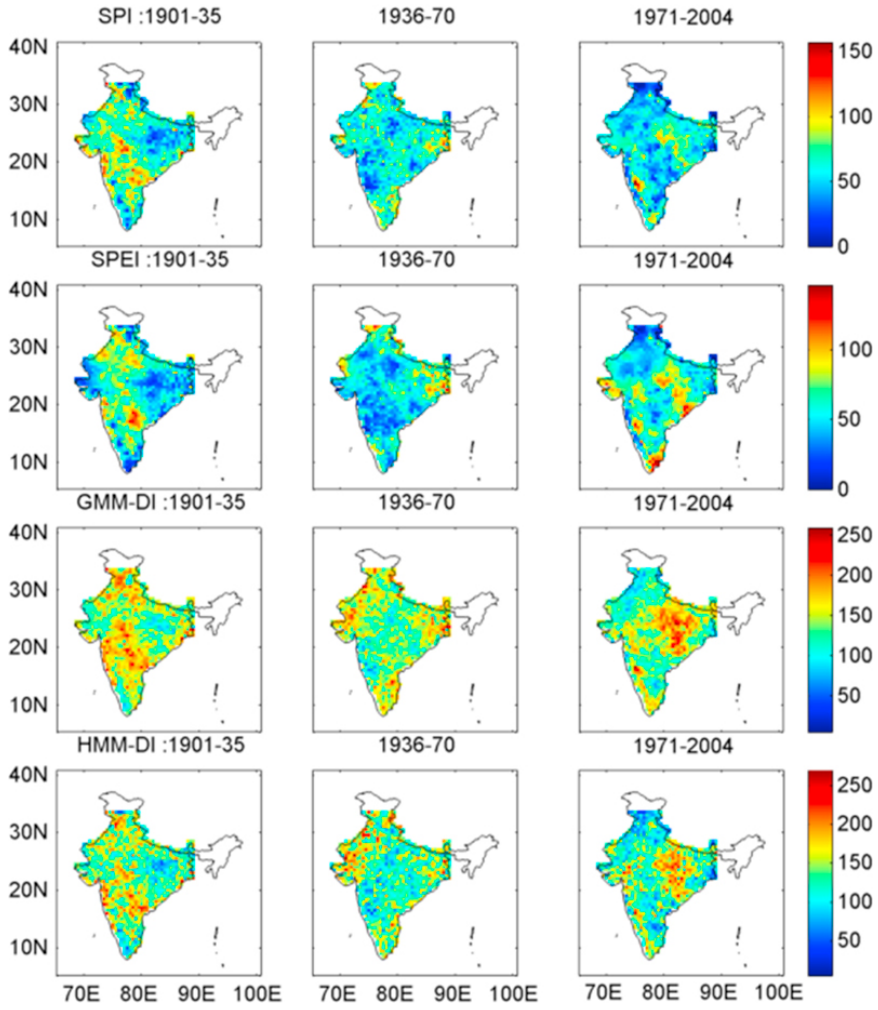

Figure 2.14. Same as Figure 2.12, but for drought duration (in months).

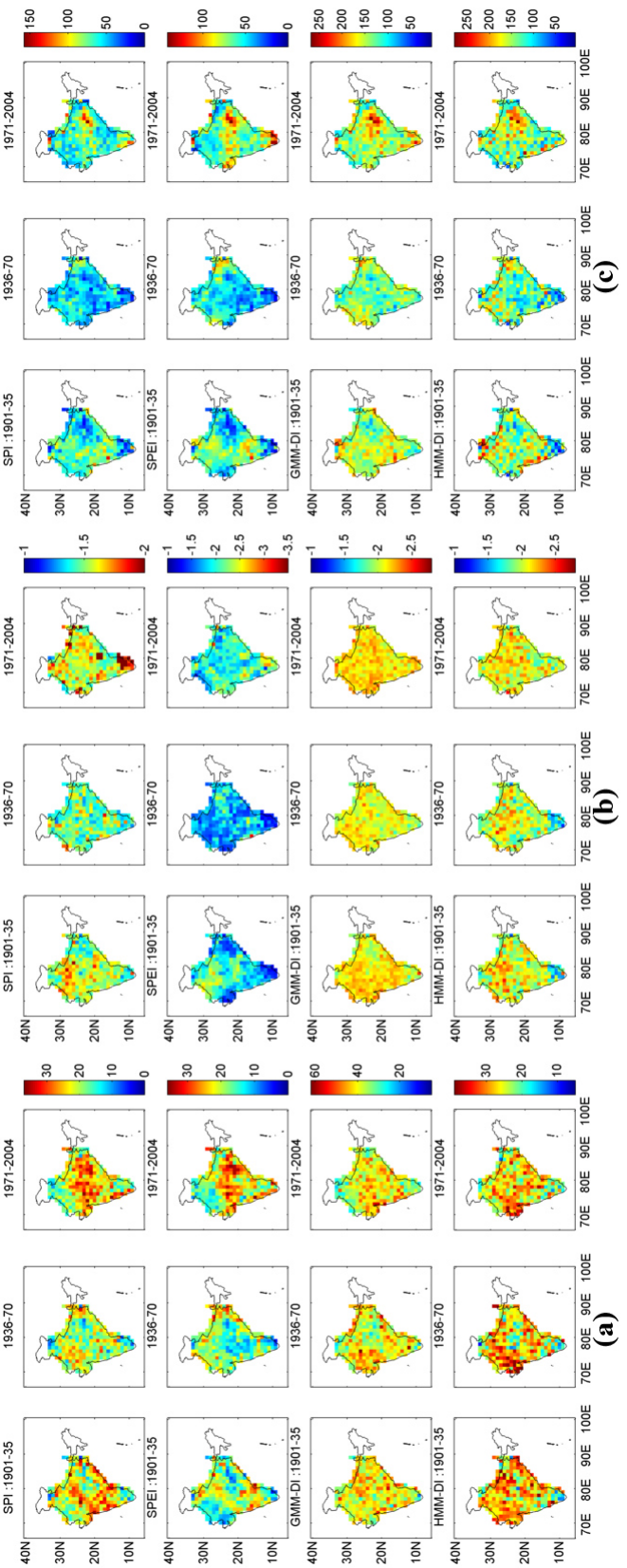

Figure 2.15. Epochal variation in 7-month drought statistics over IMR using IMD dataset where (a) Number of drought events, (b) Average intensity of drought, and (c) Duration of drought in months. In each sub-plot top panel represents SPI, followed by SPEI, GMM-DI, and HMM-DI.

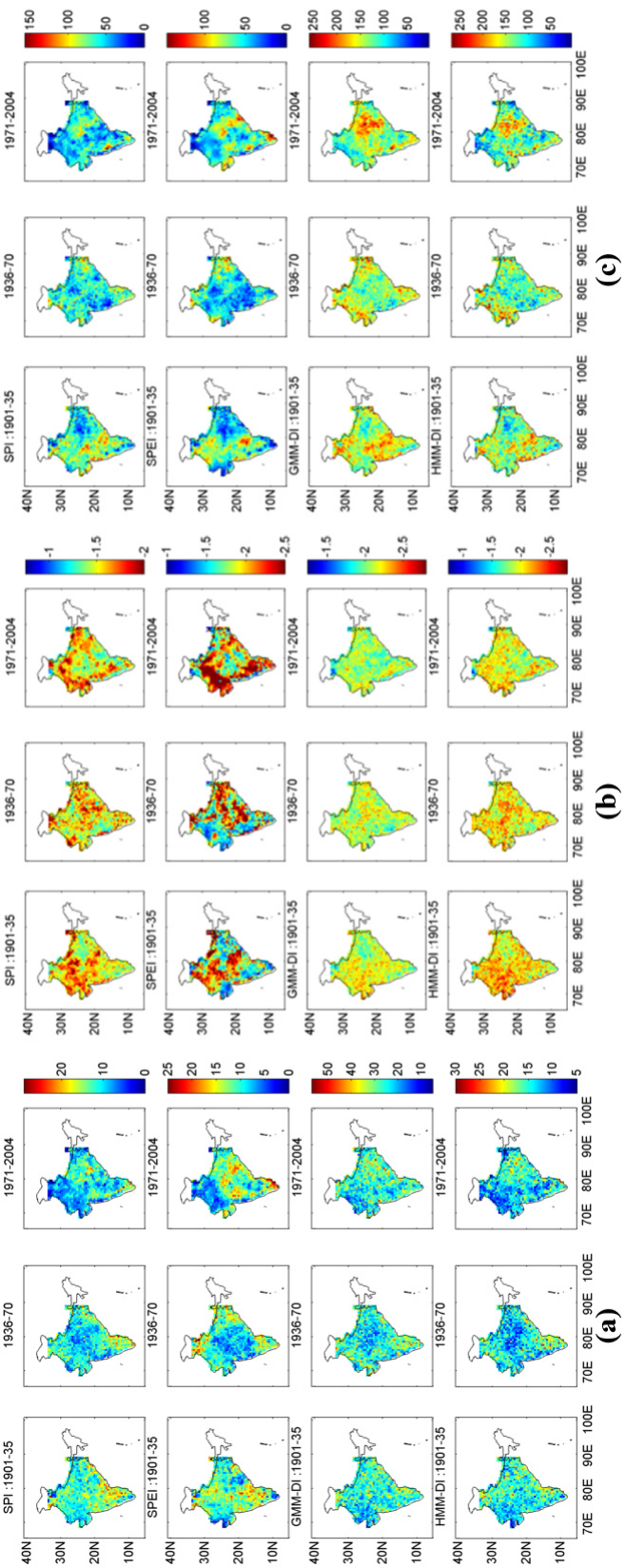

Figure 2.16. Same as Figure 2.15, but for UD dataset.

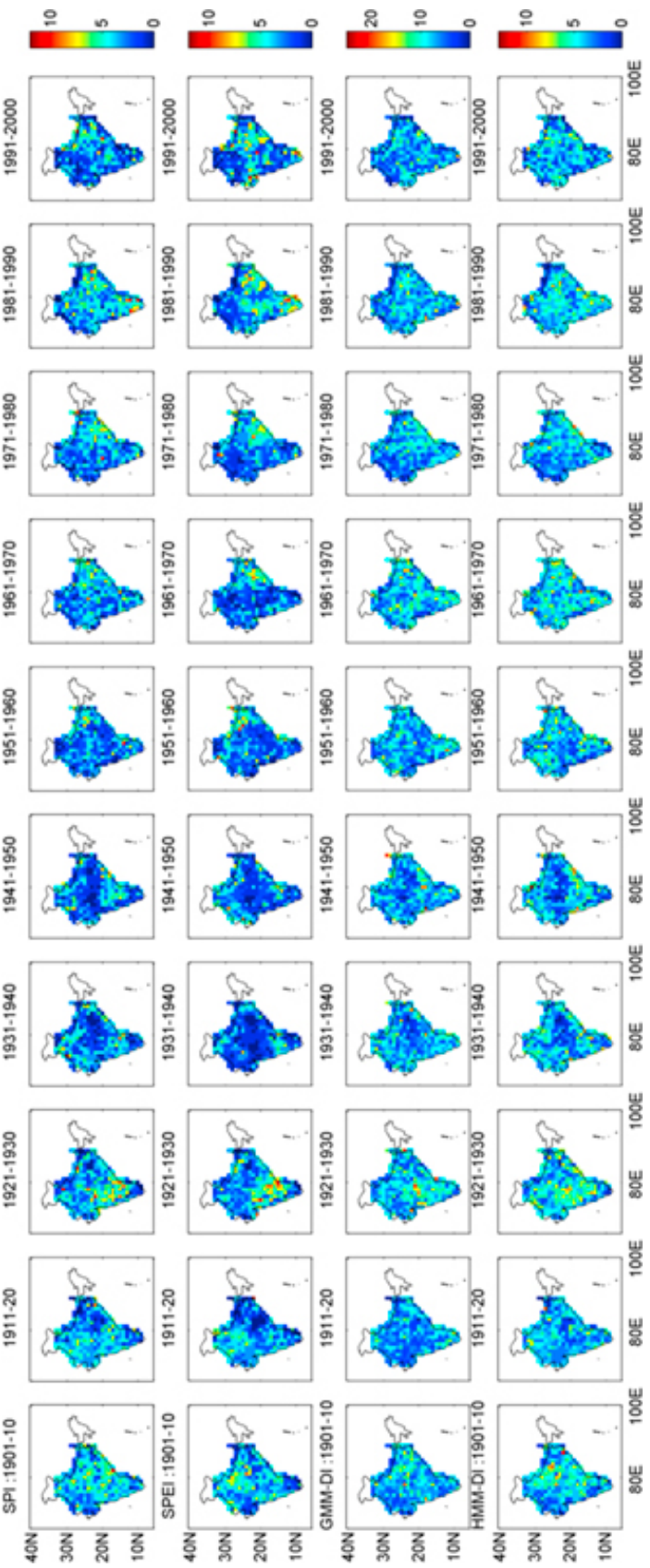

Figure 2.17. Decadal variation in number of drought events over IMR using IMD dataset. In each sub-plot top panel represents SPI, followed by SPEI, GMM-DI, and HMM-DI.

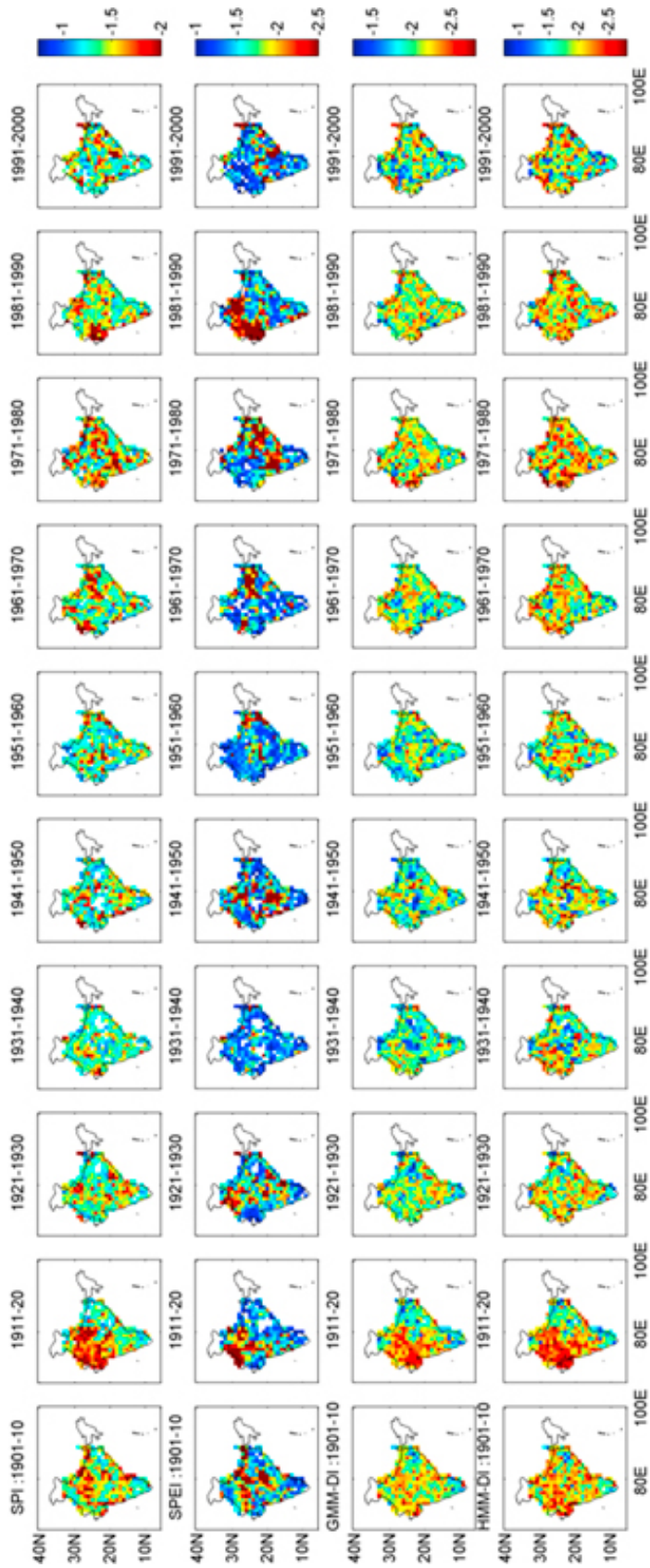

Figure 2.18. Same as Figure 2.17, but for drought intensity.

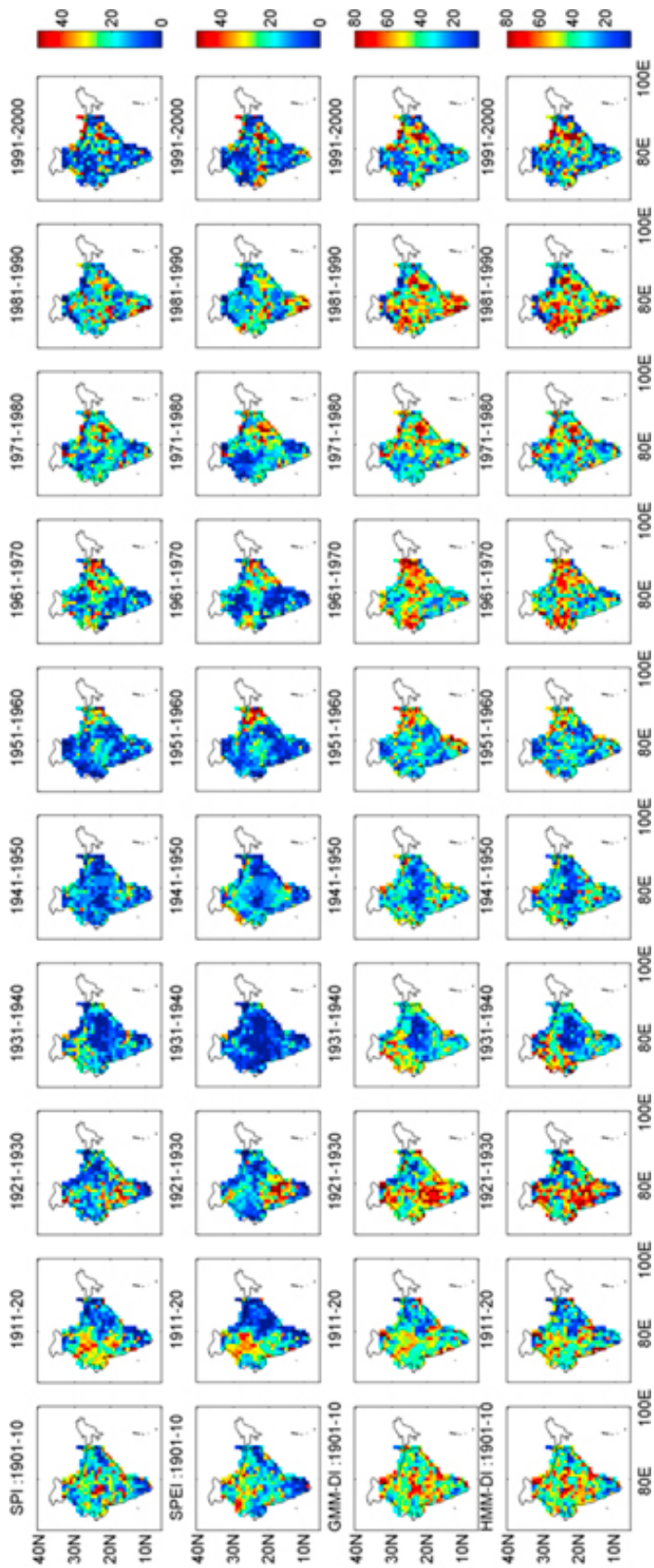

Figure 2.19. Same as Figure 2.17, but for drought duration (in months).

2.3.3 Trends

Figure 2.20 and Figure 2.21 show the trends in drought intensity computed using modified Mann-Kendall trend test, for SPI, SPEI, GMM-DI, and HMM-DI analysis for the IMD and UD datasets respectively, for 12-month time window ending in September. In the IMD dataset, for SPI analysis, during the epoch 1936-1970 (Figure 2.20a) drought intensity increased (trend is towards negative SPI values as its magnitude is negative) in the eastern IGP and parts of south-India. During the recent epoch 1971-2004, additional grids showed an increase in drought intensity in south-India (parts of coastal Tamilnadu and coastal Karnataka) and western-Rajasthan. These results are consistent with Niyogi et al. (2010) who have shown using empirical orthogonal functions and genetic algorithm-based analyses that anthropogenic land use modifications due to agricultural intensification may have resulted in significant decline in precipitation in north/northwest India and increasing patterns over east central India. Similar conclusions could be drawn from SPEI analysis (Figure 2.20b), GMM-DI analysis (Figure 2.20c) and HMM-DI analysis (Figure 2.20d). Thus parts of eastern IGP, western-Rajasthan, and parts of coastal south-India emerge as the current hotspots for droughts.

For the finer-resolution UD dataset (Figure 2.21), using a 12-month time window ending in September, each of the four drought indices shows an increasing trend in drought intensity during the period 1936-70 over the eastern IGP. However, during 1971-2004, trends in drought intensity also show an increase in south-India (parts of coastal Tamilnadu, coastal Karnataka, and central Maharashtra) and western-Rajasthan, in addition to central and eastern IGP. Thus as in case of IMD dataset, we can conclude that parts of eastern IGP, and parts of coastal south-India are emergent vulnerable regions to droughts.

At shorter time scales (e.g. 7-months ending in December) it was found that in addition to central- and eastern IGP and coastal south-India, interior parts of

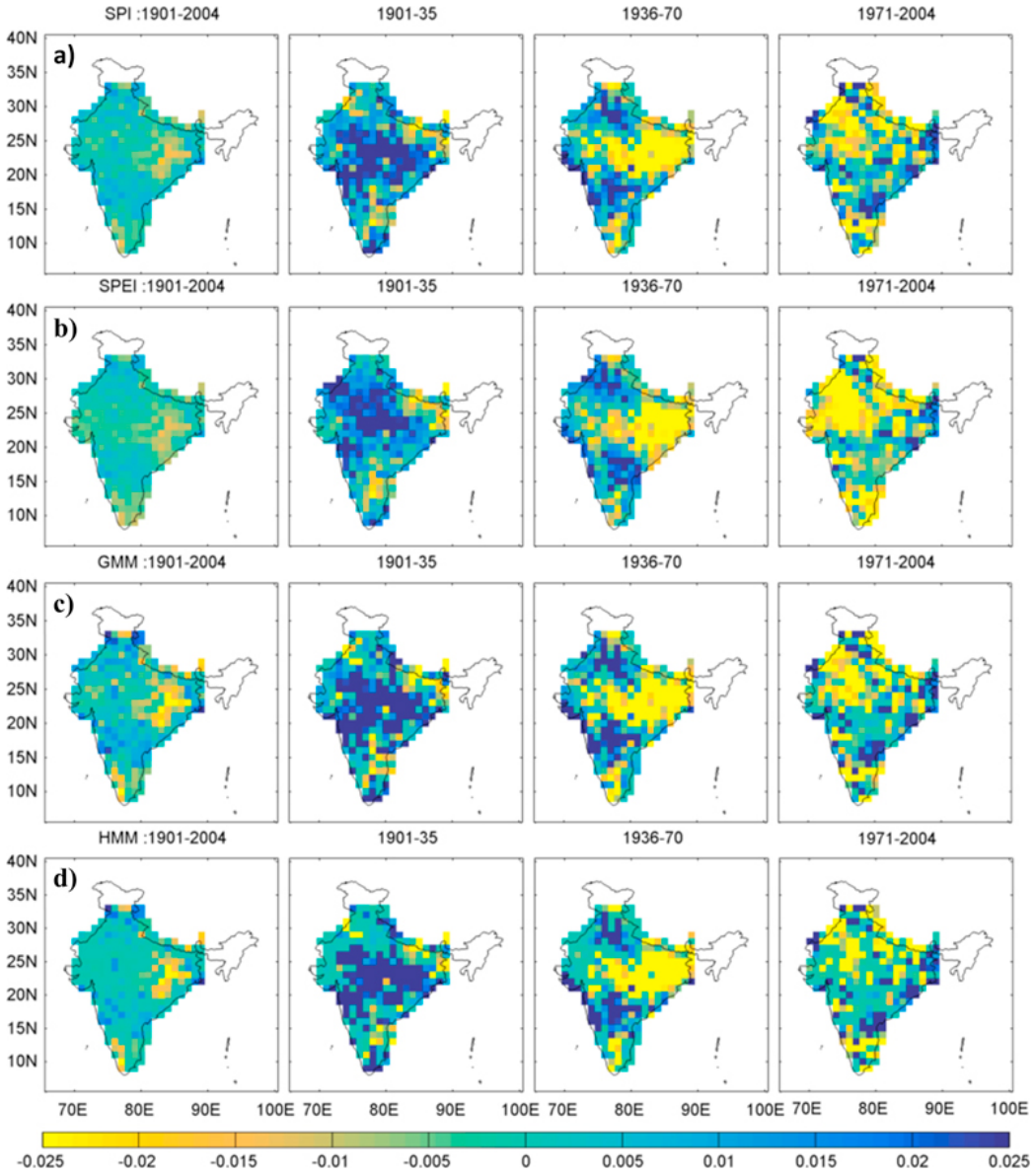

Figure 2.20. Mann-Kendall trend slope for 12-month drought intensity ending in September over IMR during the periods 1901-2004, 1902-1935, 1936-1970, and 1971-2004. Results correspond to the IMD dataset using (a) SPI, (b) SPEI, (c) GMM-DI, and (d) HMM-DI.

Maharashtra and central India were emerging as vulnerable regions to droughts for both IMD (Figure 2.22) and UD datasets (Figure 2.23).

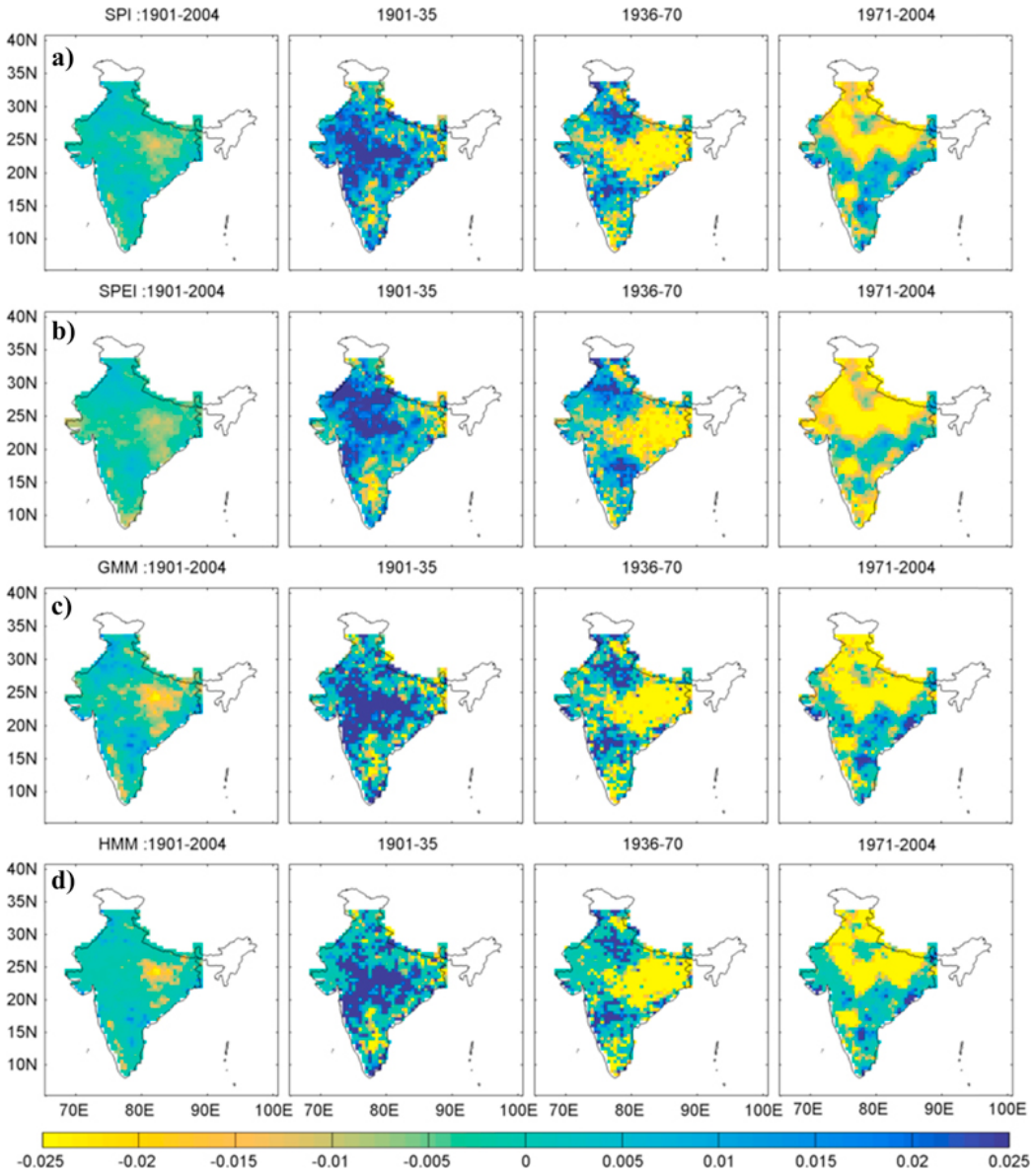

Figure 2.21. Same as Figure 2.20, but for UD.

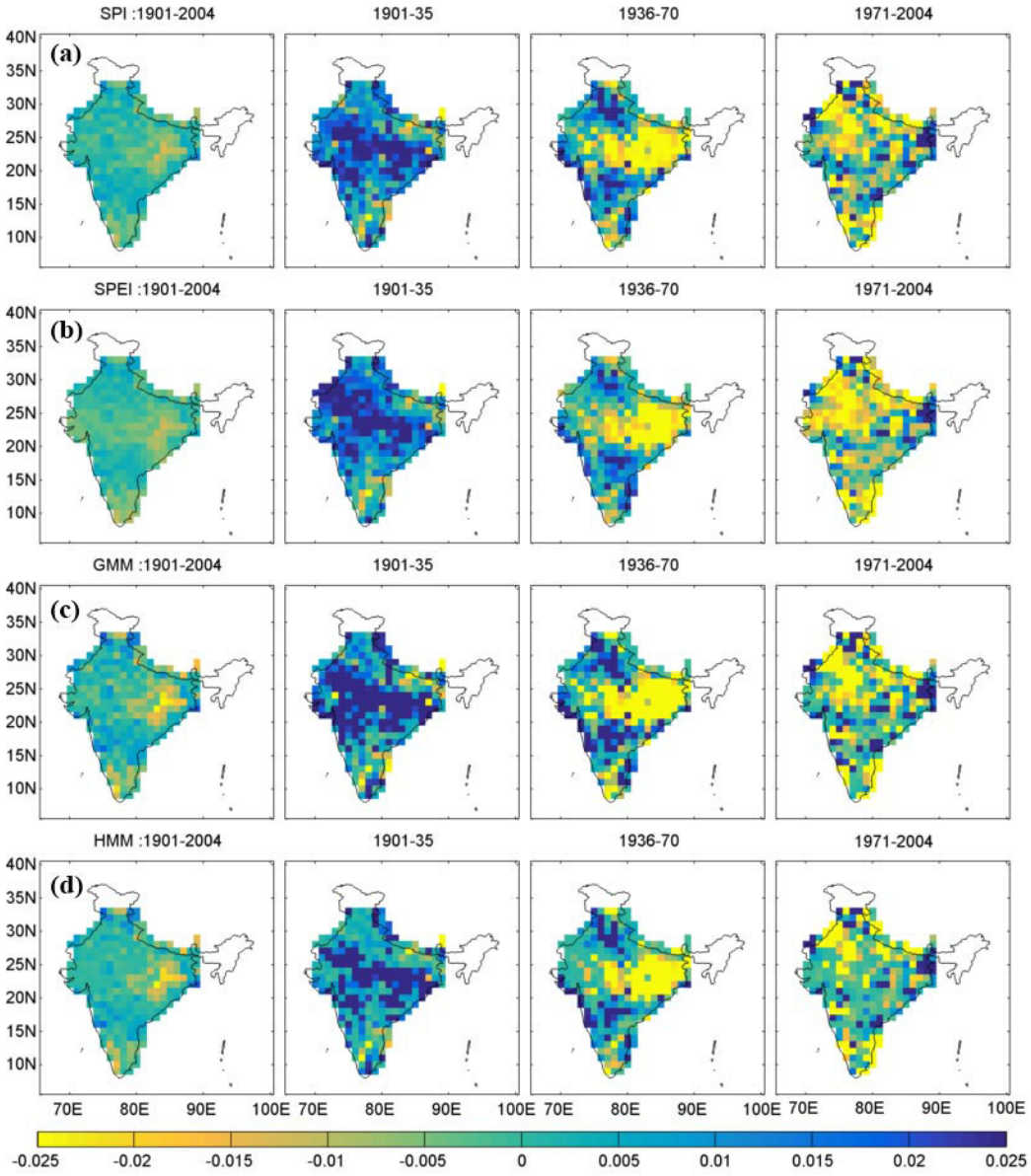

Figure 2.22. Same as Figure 2.20, but for 7-month time window ending in December.

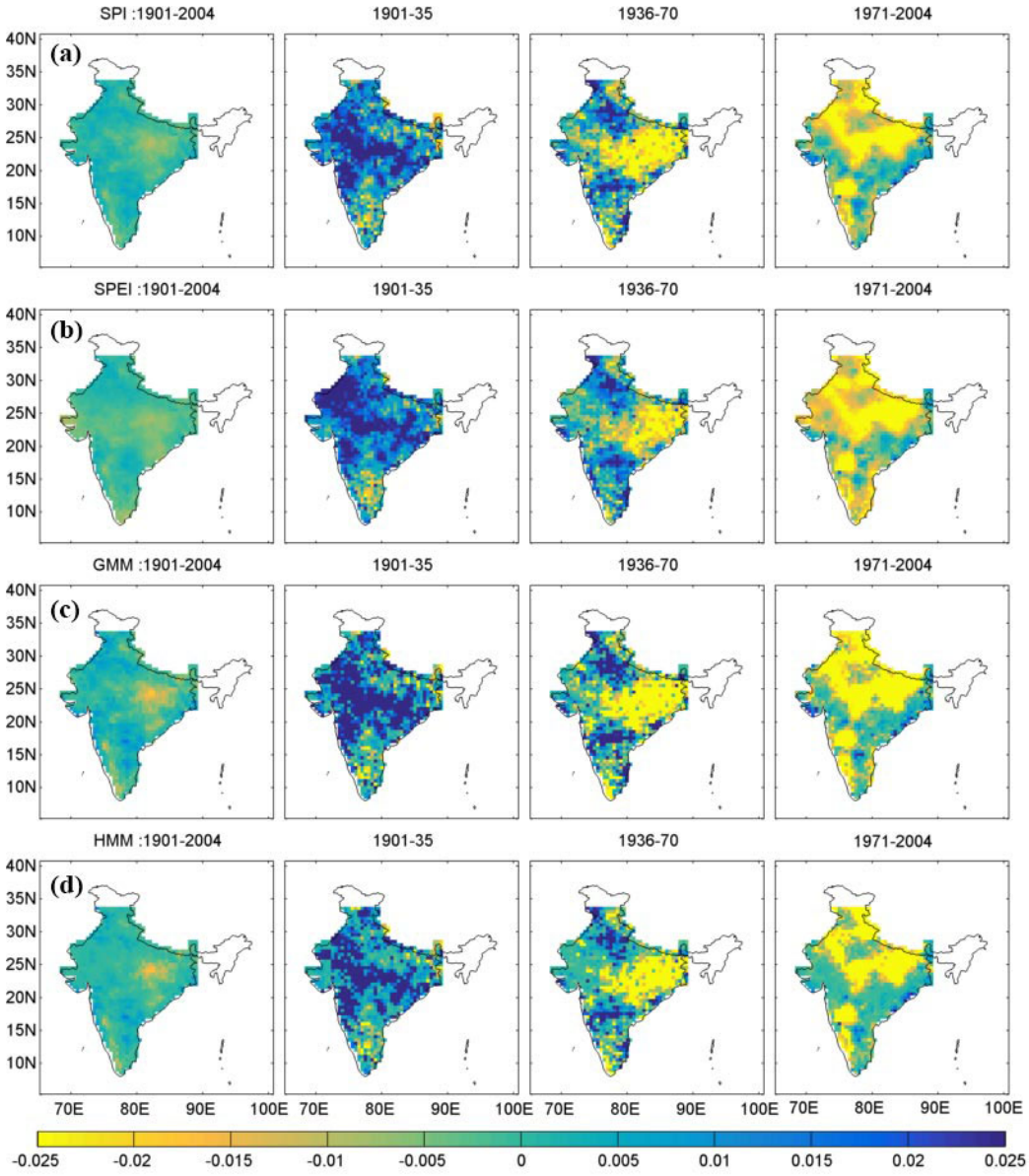

Figure 2.23. Same as Figure 2.20, but for 7-month time window ending in December using UD dataset.

2.3.4 Drought frequency

Hypothesis tests were carried out to investigate whether the number of droughts had significantly increased during the recent epoch 1971-2004, when considering 12-month droughts ending in September. A right tailed t-test with significance level of 5% was used. Figure 2.24-b show the results of the hypothesis test at each grid of the IMD and UD datasets for SPI, SPEI, GMM-DI, and HMM-DI, respectively. The results indicate that the hypothesis test was significant, or in other words the number of droughts had shown a statistically significant increase at several grids in the study region. To account for the bias induced in the hypothesis test due to spatial correlation in the gridded meteorological data, an FDR test (Ventura et al., 2004; Wilks, 2006) was performed. The FDR test provides a control for the number of falsely rejected hypothesis out of all rejected (i.e., statistically significant) hypothesis (Ventura et al., 2004). The FDR test further confirmed that the number of droughts showed a statistically significant increase in the Indo-Gangetic plains, coastal south-India, and central Maharashtra during the recent period 1971-2004.

Similarly, for 7-month time window ending in December, it was found that the number of droughts showed a statistically significant increase in the central and eastern IGP and interior parts of Maharashtra during the recent epoch (1971-2004) for both IMD (Figure 2.25a) and UD datasets (Figure 2.25b).

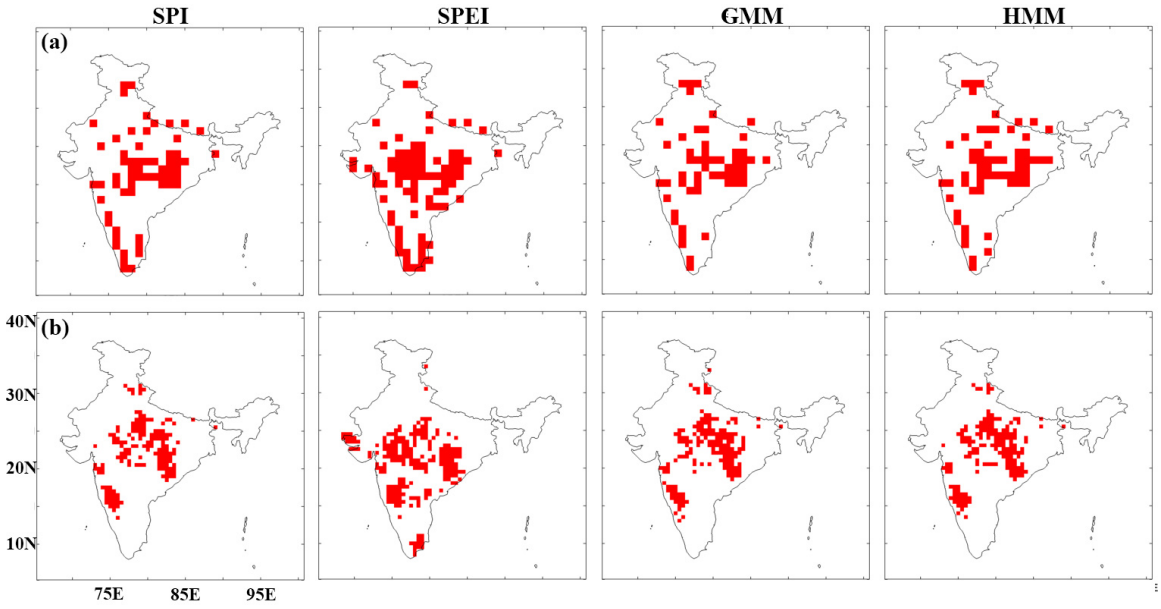

Figure 2.24. Hypothesis test to see if the number of droughts (moderate, severe and extreme) of 12-month time window ending in September have increased during the period 1971-2004 in comparison to 1936-1970 for (a) IMD and, (b) UD precipitation datasets according to SPI, SPEI, GMM-DI, and HMM-DI. Grids where the number of droughts show a statistically significant increase at $\alpha = 0.05$ are displayed.

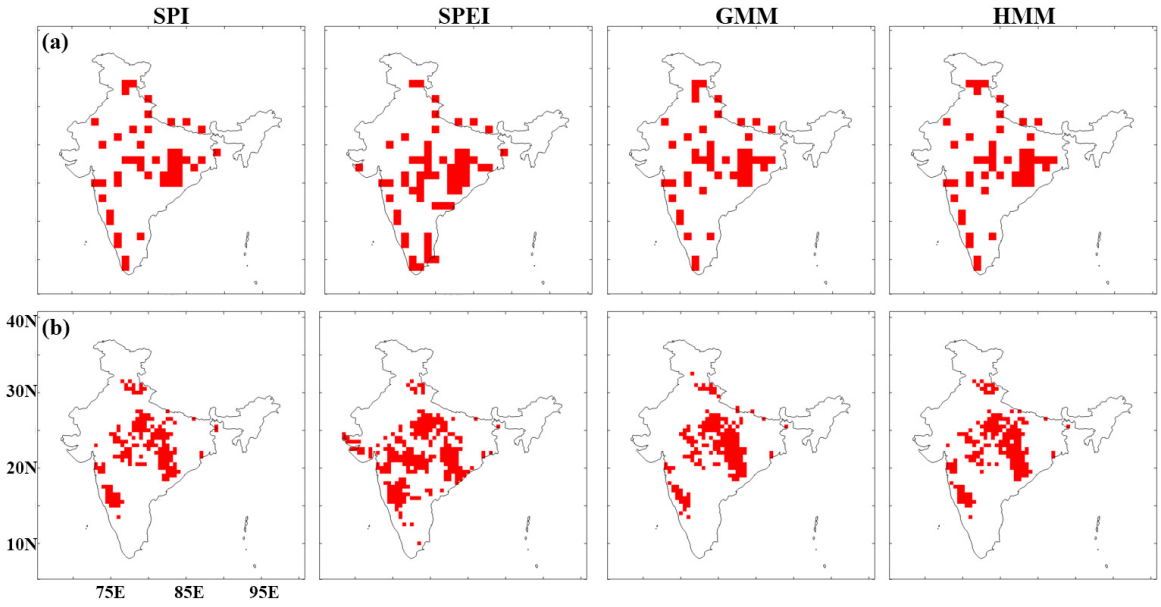

Figure 2.25. Same as 2.24, but for 7-month time window June to December.

2.3.5 Drought vulnerability

Figures 2.26a-b show the regions over IMR that were vulnerable to droughts (defined as SPI < -1.0) using IMD and UD precipitation datasets for the three study periods, considering a 12-month time window. Using the gridded population estimates available (Center for International Earth Science Information Network - CIESIN - Columbia University et al., 2005), an estimate of the population affected by droughts for the three periods was obtained. According to SPI, during the recent period of 1971-2004 approximately 405 million people were in the drought affected region. This is equivalent to a GDP of USD 208 billion. The population and GDP estimates are calculated after defining a threshold for drought intensity below which a drought is considered to have negative impact on the economy and society. The values in the bar plot (see inset in Figure 2.26a-b) correspond to an intensity threshold of -1.0 for SPI. Similar computations using SPEI, GMM-DI and HMM-DI resulted in consistently higher estimates compared to SPI for each of the three periods. This may be due to the choice of threshold and the differences in the methodology used in their computation.

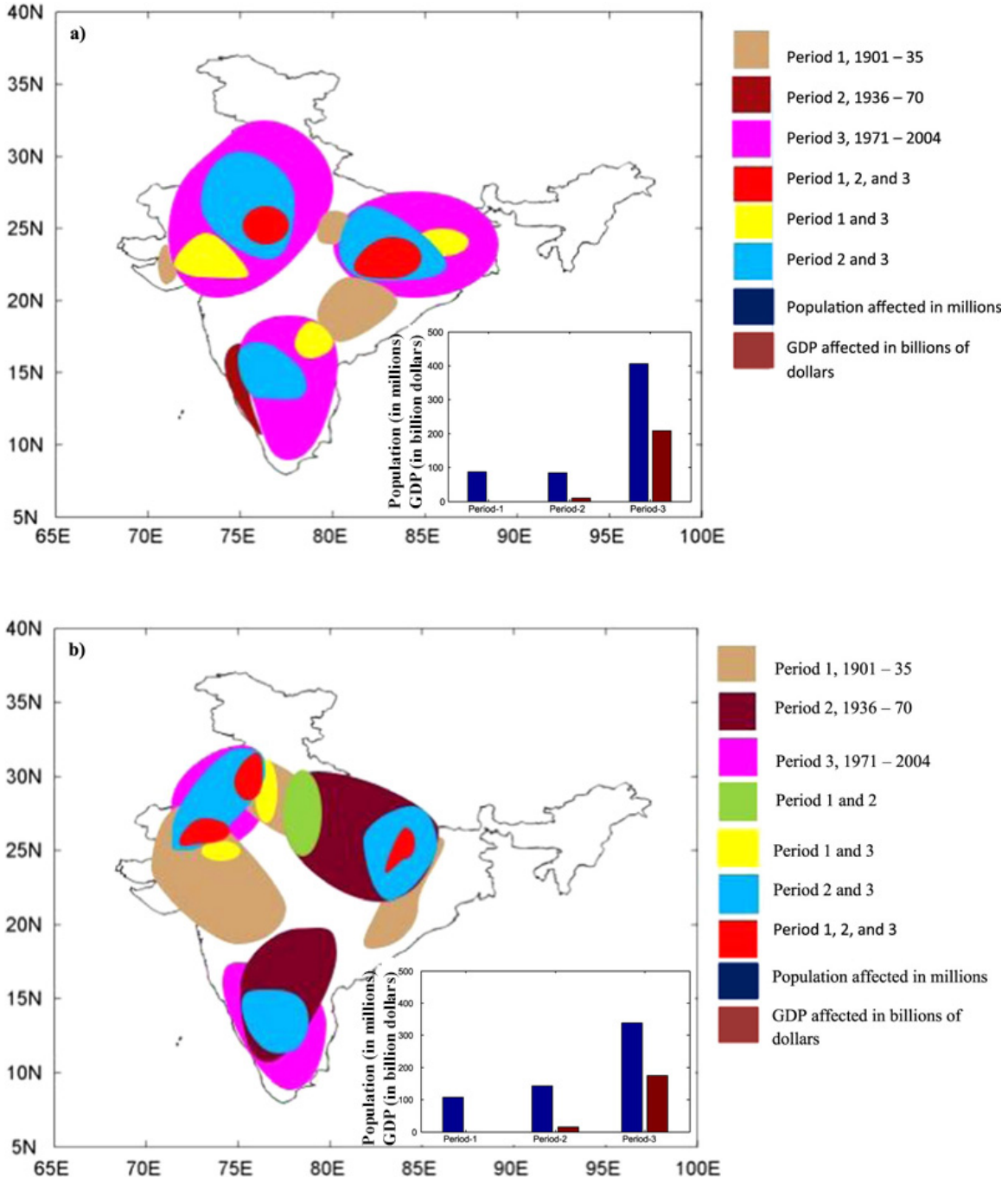

Figure 2.26. The estimate of population and GDP affected, and the drought hotspots during the sub-periods 1901-1935, 1936-1970, and 1971-2004 according to SPI< -1.0 for (a) IMD precipitation dataset and (b) UD precipitation dataset.

2.4 Summary and concluding remarks

Recent studies have highlighted that IMR has a steady increase in the drought patterns. Motivated by the cautionary conclusions of Trenberth et al. (2014), a reassessment of the drought patterns using multiple data sources and methods was conducted. Accordingly, long-term retrospective drought variability over the IMR was examined using two gridded precipitation datasets that differ in their primary data source and spatial resolution. Moreover, several drought characteristics (severity, duration, areal extent, and frequency) were compared using SPI, SPEI, GMM-DI, and HMM-DI to assess the variability in the results.

SPI, SPEI, GMM-DI, and HMM-DI were analyzed for three periods 1901-1935, 1936-1970, and 1971-2004 to examine epochal and decadal variation in drought characteristics over the IMR. Consistent with the findings from recent studies that indicate monsoon precipitation is becoming extreme and regionally varied, significant change in the drought climatology over the IMR was noted. Results indicated that droughts were becoming much more regional in recent decades and showing a general migration from west to east in the Indo-Gangetic plain. An increased duration, severity, and spatial extent in recent decades were observed, and the Indo-Gangetic plain, parts of coastal south-India and central Maharashtra were identified as vulnerable regions for recent droughts. Despite some differences in results for the choice of drought indices, the time window chosen for analysis, and/or the precipitation dataset (resolution) used, overall the results and conclusions are consistent.

It is beyond the scope of present study to assess the causal mechanism of droughts, and to investigate if the observed trends are related to other phenomena such as changes observed in the monsoon break (active - dry spell) periods (Singh et al., 2014). There are a number of possible mechanisms - aerosols, landuse change, SST changes, global changes, thermodynamic feedback due to heating rates (Roxy et al., 2015); as a result, diagnosis and discussion of potential mechanisms will have to be a part of follow up study. The results from this study provide the baseline for

future climate change studies, and yield the robust conclusion that irrespective of the datasets and methodology used, the IMR has high potential for droughts, and that droughts appear to be migrating to the agriculturally important regions including Indo Gangetic plains.

3. PROBABILISTIC DROUGHT CLASSIFICATION WITH STANDARDIZED PRECIPITATION INDEX

This article has been previously published in Journal of Hydrology July 2015.

3.1 Introduction

Drought classification schemes classify a drought based on its severity or intensity. Water resources planners rely on drought classification to decide on 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 USA by classifying droughts into five classes - D0 to D4 with the latter representing an 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-term 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. (2013) 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 have been addressed in this work.

Drought classification schemes employ drought indices that measure the 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 relationships 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 Jurez, 2001; Ryu et al., 2010). Among several drought indices proposed in the literature (Dai, 2011; Heim, 2002; Mishra and Singh, 2010), the SPI (McKee et al., 1993) is very popular because of its computational simplicity and versatility in comparing different 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 how 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 estimates providing discrete classification. Thus, 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 the data. McKee et al. (1995) in their original paper on SPI recommend the gamma distribution. Lloyd-Huges and Saunders (2002) found the 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) argue that different distributions may be appropriate for different drought durations (window size), and recommend 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. (2013) used 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 for model uncertainties; however, the number of hidden states (drought classes) was pre-specified. To facilitate comparison of HMM-based 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 the over-specification problem, i.e. the model structure is more complicated than can be supported by the dataset. Specifically, in the HMM context, over-specification implies that the number of specified hidden states are more than that needed to model the data. Over-specification can result in parameter identification problems leading to unreliable results.

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

tic 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 chapter 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.

3.2 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). Figure

3.1 shows the study area along with the grid locations for which rainfall data were available.

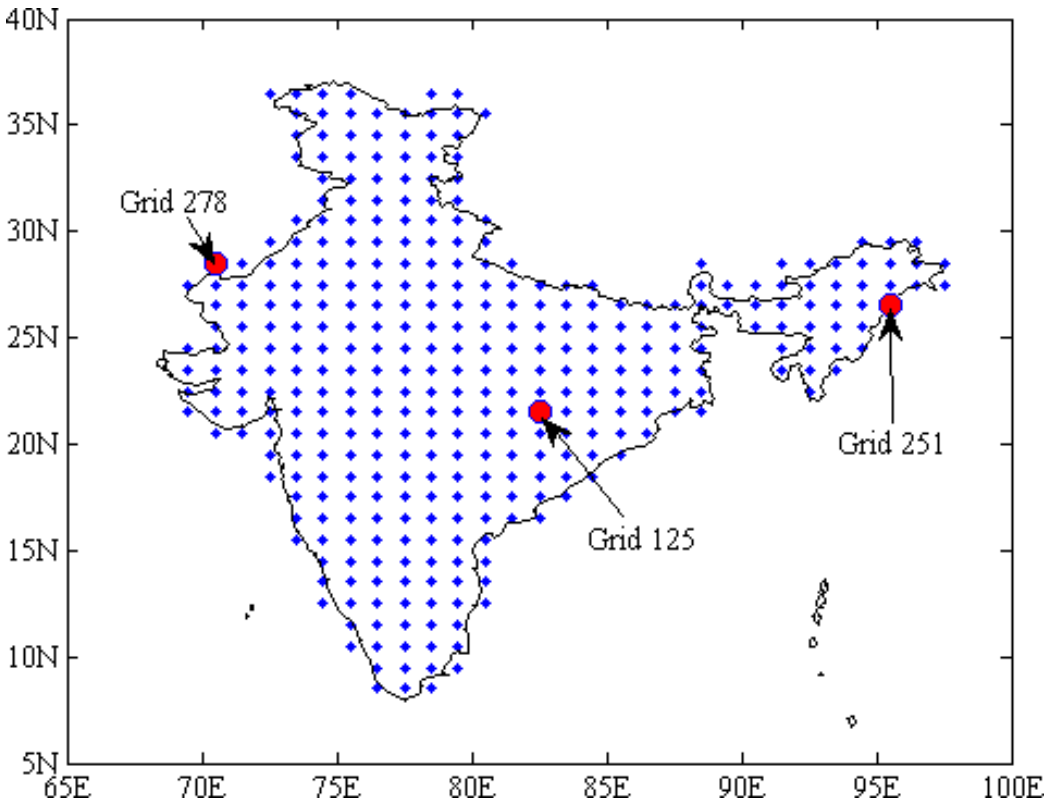

Figure 3.1. Map showing the study area along with the location of grids for which rainfall data were provided by IMD.

3.3 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.

3. For a given rainfall event, determine the 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. Thresholds on the CDF function should be decided based on the application of the drought classification scheme. To draw parallels with the US Drought Monitor, the same thresholds as used by them for SPI have been adopted for drought classification (Table 3.1).

Table 3.1.
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 to 0.309
D1	Moderate Drought	-0.8 to -1.3	0.097 to 0.212
D2	Severe Drought	-1.3 to -1.6	0.055 to 0.097
D3	Extreme Drought	-1.6 to -1.9	0.023 to 0.055
D4	Exceptional Drought	-2.0 or less	0.023 or less

3.4 Gamma mixture model

As discussed in the Introduction section (3.1), there is an ongoing debate on the choice of a suitable distribution for fitting data in SPI analysis. This problem is addressed 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, a brief description of the Gamma-MM is provided. Readers are referred to Wiper et al. (2001) and Richardson and Green (1997) for details on mixture models.

Let the cumulative precipitation 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,

$$p(x_t|\lambda) = \sum_{i=1}^{M} w_i G\left(x_t|\nu_i, \frac{\nu_i}{\mu_i}\right) \tag{3.1}$$

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

$$G\left(x_t|\nu_i, \frac{\nu_i}{\mu_i}\right) = \frac{\left(\frac{\nu_i}{\mu_i}\right)^{\nu_i}}{\Gamma(\nu_i)} x_t^{(\nu_i - 1)} \exp\left(-\frac{\nu_i}{\mu_i} x_t\right) \tag{3.2}$$

with mean μ_i and shape parameter ν_i. Further, the mixture weights satisfy the constraint $\sum_{i=1}^{M} w_i = 1$. The parameter set is represented as, $\lambda = \{w, \mu, \nu\}$, where $w = [w_1, w_2, \ldots, w_M]^T,\ \mu = [\mu_1, \mu_2, \ldots, \mu_M]^T$ and $\nu = [\nu_1, \nu_2, \ldots, \nu_M]^T$.

In the Bayesian framework, the model parameters are obtained by specifying prior distributions to model parameters. The parameter estimation can be simplified by introducing a latent variable $Z = [z_1, z_2, \ldots, z_N]^T$ for each time step. The variable z_t is an M-dimensional binary random variable, $z_t = [z_{t1}, z_{t2}, \ldots, z_{tM}]^T$, 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 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 z_t is

$$p(x_t|z_{ti} = 1) \sim\ G\left(x_t|\nu_i, \frac{\nu_i}{\mu_i}\right) \tag{3.3}$$

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

$$p(\lambda|X) \propto p(X|\lambda)p(\lambda) \tag{3.4}$$

where the parameter set λ includes the latent variable as well. The *likelihood function* given the latent variable is

$$p(X|\lambda) = p(X|Z,\mu,\nu) = \prod_{t=1}^{N}\prod_{i=1}^{M}\left(G\left(x_t|\nu_i, \frac{\nu_i}{\mu_i}\right)\right)^{z_{ti}}$$

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

$p(\lambda) = p(Z|w)p(w)p(\mu)p(\nu)$ with

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

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

$$p(\nu) = Exp(\nu|\theta) = \prod_{i=1}^{M}\frac{1}{\theta_i}\exp(-\theta_i\nu_i),\ \theta = [\theta_1, \ldots, \theta_M]^T, \text{ and}$$

$$p(\mu) = GI(\mu|\alpha,\beta) = \prod_{i=1}^{M}\frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)}\mu_i^{-\alpha_i-1}\exp(-\frac{\beta_i}{\mu_i}),\ \alpha = [\alpha_1, \ldots, \alpha_M]^T \text{ and } \beta = [\beta_1, \ldots, \beta_M]^T$$

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 or Markov chain Monte Carlo methods). In this study the posterior distribution is estimated using stochastic approximation by sampling the posterior distribution with Gibbs sampler, an MCMC algorithm (Geman and Geman, 1984) as described in Section 3.4.1.

In the above formulation of Gamma-MM, it is assumed that the number of mixture components, M, is known. However, in a 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. This study employs 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. An alternative approach is to 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. The latter approach is recommended for hydrological applications where the number of components is usually limited to 2 or 3.

In the Bayesian framework, mixture models frequently suffer from an *identifiability* problem i.e., a M component mixture model will have a total of $M!$ equivalent solutions. The problem can be avoided by introducing asymmetry 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).

3.4.1 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 follows:

1. Set iteration number $j = 0$, and parameters to their initial value $\lambda^{(0)} = [w^{(0)}, \mu^{(0)}, \nu^{(0)}]$. The initial value is obtained by randomly sampling from the prior distribution of the parameters.

2. Sample from $p\left(z_t^{j+1}|X, w^{(j)}, \mu^{(j)}, \nu^{(j)}\right) \sim \ Multinomial(z_t|r_t)$
   where $r_t = [r_{t1}, \ldots, r_{tM}]^T$, $r_{ti} = \dfrac{S_{ti}}{\sum_{i=1}^{M} S_{ti}}$ and $s_{ti} = w_i G\left(x_t|\nu_i, \frac{\nu_i}{\mu_i}\right)$ and $Multinomial$ represents multinomial distribution.

3. Sample from $p\left(w^{(j+1)}|X, Z^{(j+1)}, \mu^{(j)}, \nu^{(j)}\right) \sim \ Dir(w|\Phi)$
   where $\hat{\Phi} = [\phi_i + n_i, \ldots, \phi_M + n_M]^T$ and $n_i = \sum_{t=1}^{N} z_{ti}$.

4. Sample from $p\left(\mu^{(j+1)}|X, Z^{(j+1)}, w^{(j+1)}, \nu^{(j)}\right) \sim \ GI(\mu|\hat{\alpha}, \hat{\beta})$
   where $\hat{\alpha} = [\alpha_i + n_i\nu_i, \ldots, \alpha_M + n_M\nu_M]^T$ and
   $\hat{\beta} = \left[\beta_i + \nu_i \sum_{t=1}^{N} x_t z_{ti}, \ldots, \beta_M + \nu_M \sum_{t=1}^{N} x_t z_{tM}\right]^T$.

5. Sample from $p\left(\nu^{(j+1)}|X, Z^{(j+1)}, w^{(j+1)}, \mu^{(j+1)}\right)$. 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 $p(\tilde{\nu_i}|\nu_i) \sim \ G(h, h|\nu_i)$ and is accepted with a probability $min\left\{1, \frac{f(\tilde{\nu_i})p(\nu_i|\tilde{\nu_i})}{f(\nu_i)p(\tilde{\nu_i}|\nu_i)}\right\}$
   where $f(\nu_i) \propto \dfrac{\nu_i^{n_i\nu_i}}{\Gamma(\nu_i)^{n_i}} \exp\left(-\nu_i\left(\theta_i + \dfrac{\sum_t x_t z_{ti}}{\mu_i} + n_i log\mu_i - log\left(\prod_{t=1;z_{ti}=1}^{N} x_t\right)\right)\right)$
   If the new sample $\tilde{\nu_i}$ is rejected, the current value of ν_i is retained. The above procedure is repeated to sample ν_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, 15000 samples were generated after ignoring an initial 500 samples (*burn-in* period). Trace plots of the samples were monitored for convergence.

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

3.5 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 (Section 3.3), first, an annual time-series of cumulative rainfall during the monsoon season and water-year is computed. Next, the droughts are classified applying the traditional SPI and the proposed approach. Both 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 Figure 3.1) 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 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.

3.5.1 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. Figure 3.2 shows the empirical cumulative 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 the two-parameter gamma distribution.

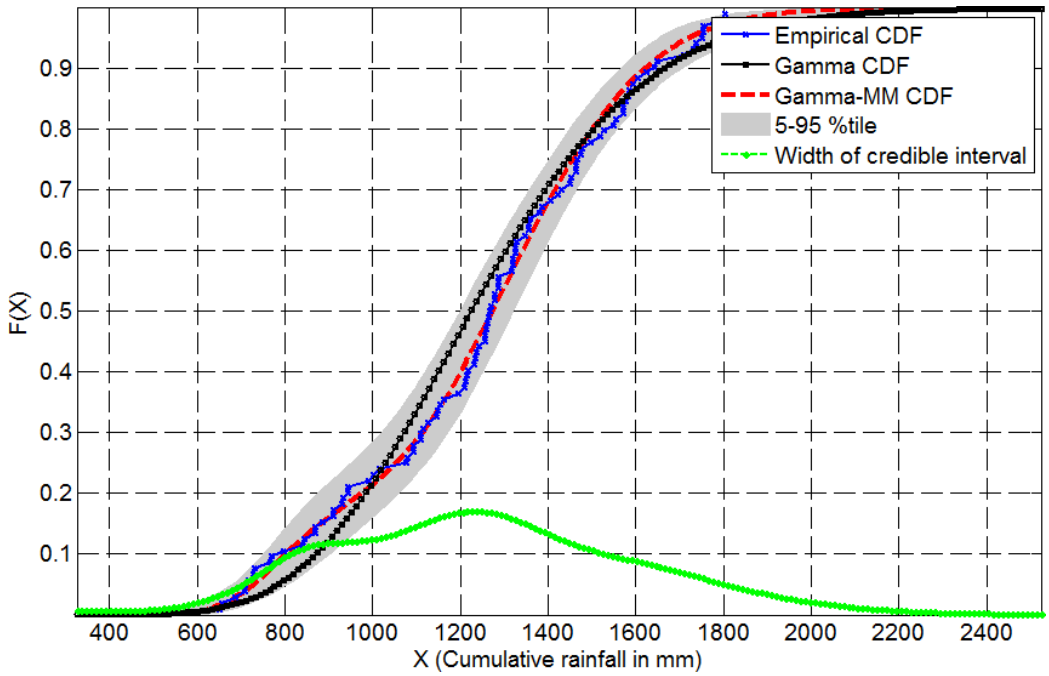

Figure 3.2. Empirical CDF along with CDFs obtained by fitting gamma distribution (Gamma CDF) and gamma mixture model (Gamma-MM CDF) to the cumulative rainfall in a water-year at Grid 125. The grey band shows 5^{th} and 95^{th} percentile of the Gamma-MM CDF and the green dotted line shows the width of its credible interval.

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. Figure 3.3 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.

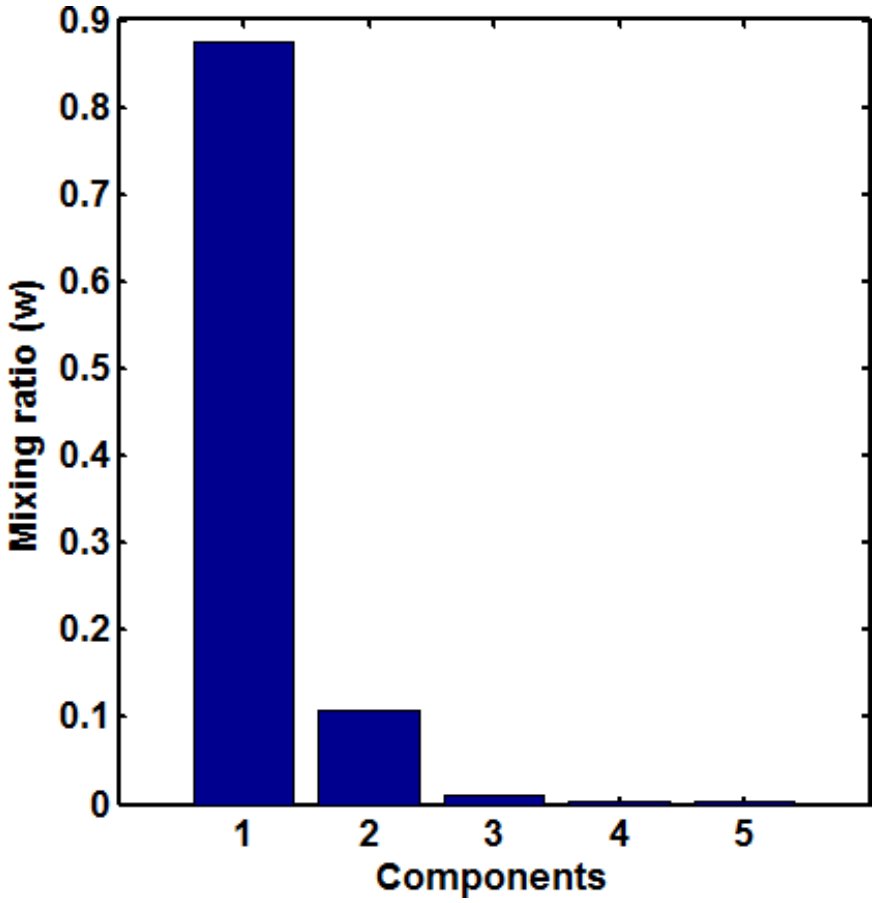

Figure 3.3. Mixing ratios of the components of a Bayesian Gamma-MM. Two components are identified to be significant for characterizing water-year drought at Grid 125.

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 and the CDF is obtained. Unlike the maximum likelihood approach that yields a point estimate of CDF, the Bayesian approach treats the CDF as a random variable and yields the distribution of CDFs for a given value of rainfall. The grey shaded band in Figure 3.2 represents 90% credible interval (5^{th} and 95^{th} percentile). The credible interval's width represents the uncertainties in estimated parameters based on available data and therefore represents epistemic uncertainty. 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; Figure 3.5), and a monotonic decreasing trend on either side of the median. Figure 3.5 shows that the CDF curve becomes almost horizontal, meaning a small change in the value of the CDF yields a large variation in rainfall values. Further, the credible interval is smaller for extreme rainfall events than wider intervals for mid-range rainfall events. However, when the credible interval is normalized with respect to the PDF (see Figure 3.4), the uncertainty associated with extreme rainfall events is revealed. Typically, if more data (longer records) were available, the credible interval in the middle portion of the CDF would reduce compared to the extreme region as there are only a smaller number of extreme events.

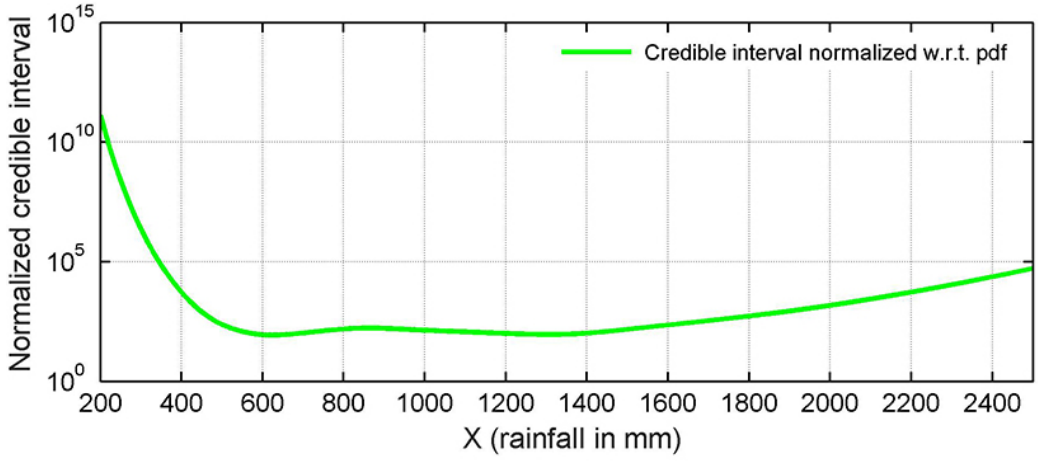

Figure 3.4. Normalized credible interval for Gamma-MM model at Grid 125 shown in Figure 3.2.

The width of the credible interval is large even for smaller values of CDFs that decide drought classes in SPI methodology. This study used credible interval of CDF for drought classification. Figure 3.6(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

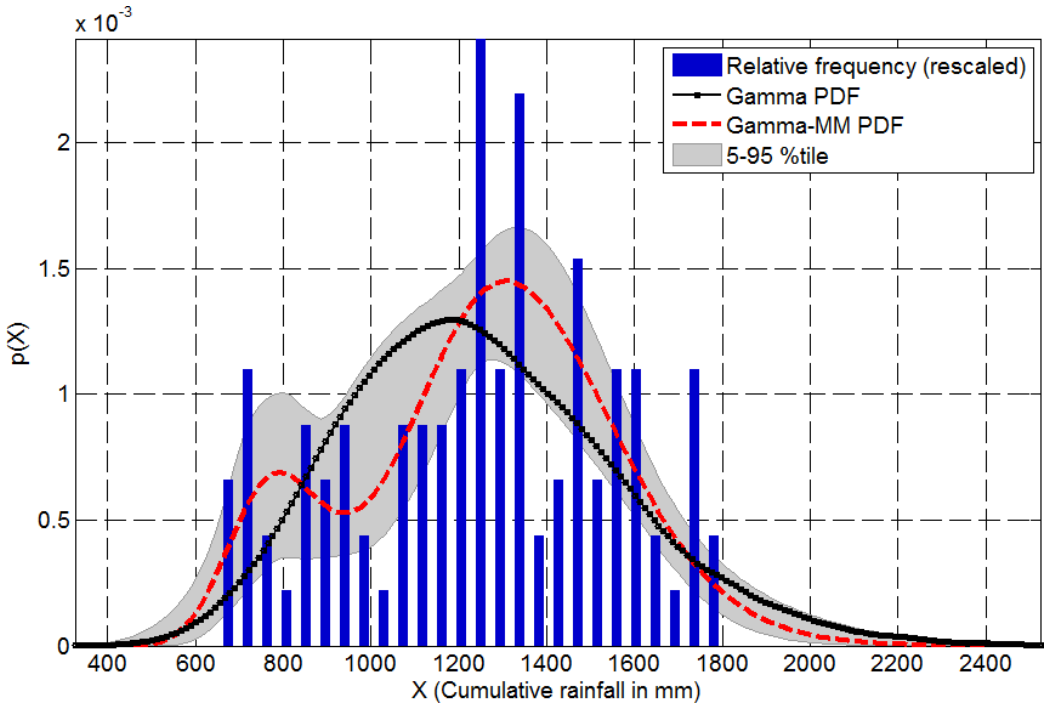

Figure 3.5. Relative frequency of the cumulative rainfall amounts in a water-year at Grid 125, and probability density functions of the fitted gamma distribution (Gamma PDF) and gamma mixture model (Gamma-MM PDF). The grey band shows 90% credible interval (5^{th} and 95^{th} percentile) of the Gamma-MM PDF.

thresholds, hence the boundaries separating two drought classes are vertical lines on the panel. The top panel of Figure 3.6 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 the 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 where non-exceedance probability of the cumulative rainfall is less than 0.023 [$F(X) < 0.023$; Table 3.1], i.e. the magnitude of rainfall during a D4 drought event is less than the amount of rainfall having the non-exceedance

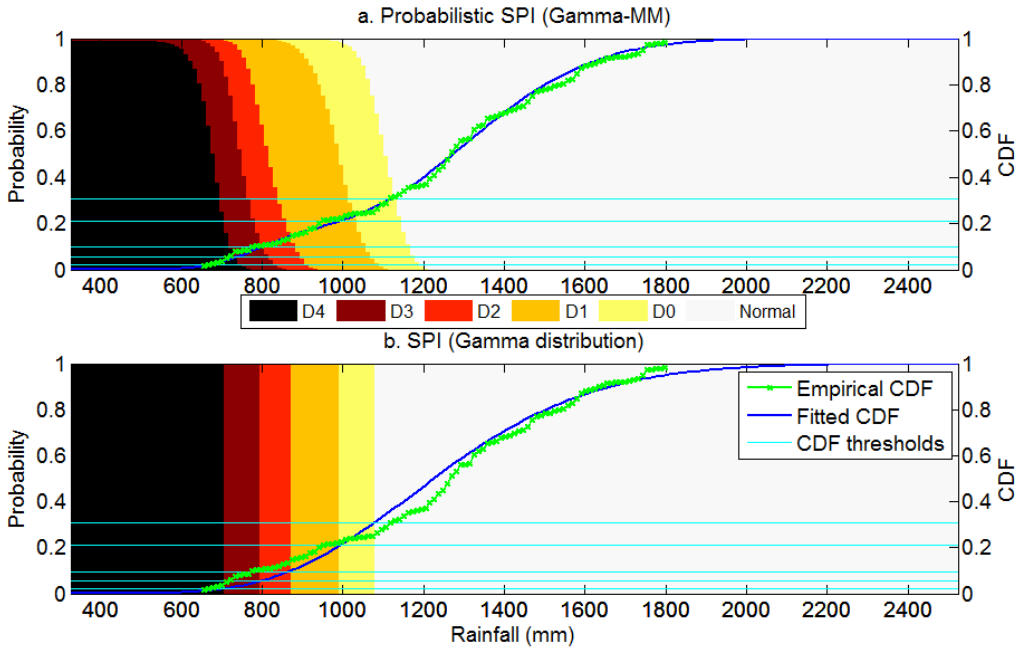

Figure 3.6. Drought classification using rainfall at Grid 125 by the probabilistic SPI (top panel) and standard SPI (bottom panel). The colored patches represent drought classes, the light horizontal lines denote thresholds on CDF specified by US Drought Monitor, and the solid curves represent empirical and fitted CDFs.

probability of 2.3%. 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.

Figure 3.7 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 (Figure 3.7a) and standard SPI methods (Figure 3.7c) 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 categorised 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 Figure 3.7b. Compared to the probabilistic SPI results (Figure 3.7a), 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.

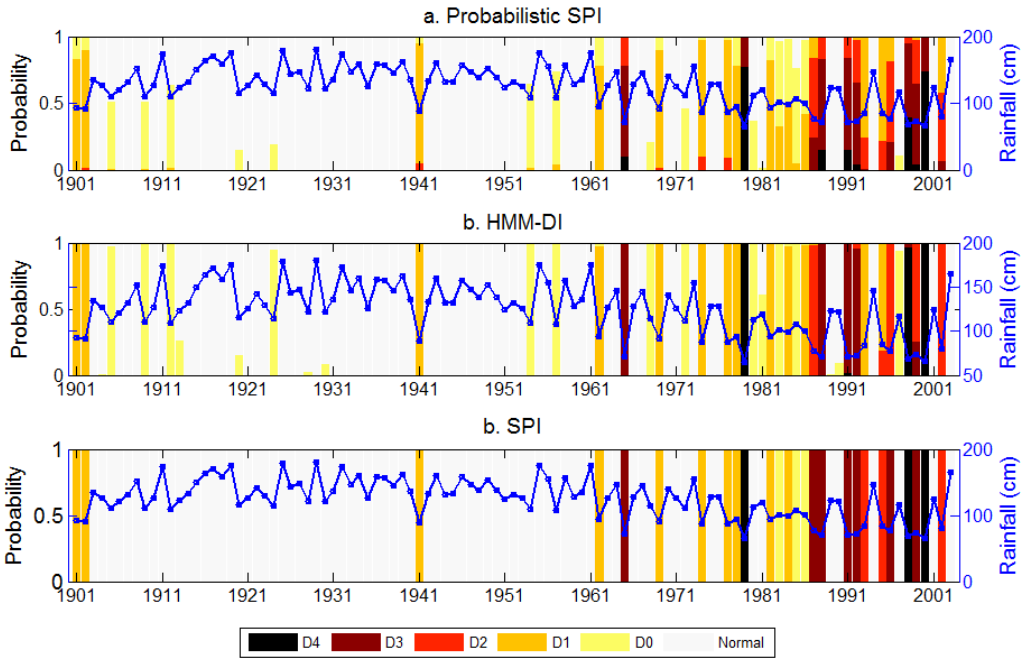

Figure 3.7. Classification of historical droughts during a water-year at Grid 125 using probabilistic SPI, HMM-DI, and standard SPI approaches. The solid blue line represents cumulative rainfall during a water-year, a colored bar denotes drought classes and its length represents probability of drought state.

Figure 3.8 shows the relative frequency of the rainfall during the monsoon months, JJAS, at Grid 125. As in the case of water-year rainfall (Figure 3.5), the monsoon rainfall also exhibits two distinct modes that are captured by the 2-component Gamma-MM but missed by the gamma distribution. Figure 3.9 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, Figure 3.8) and a monotonic decreasing trend away from the median, similar in nature to Figure 3.5. Figure 3.10 presents the demarcating boundaries for the drought classes determined by the two methods. As in the case of water-year droughts (Figure 3.6), 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 (Figure 3.11). 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 (5cm and 2cm, respectively). Probabilistic SPI classifies all the three years in D0 and D1 classes with probabilities 60% & 39%, 19% & 81%, and 37% & 63%, respectively.

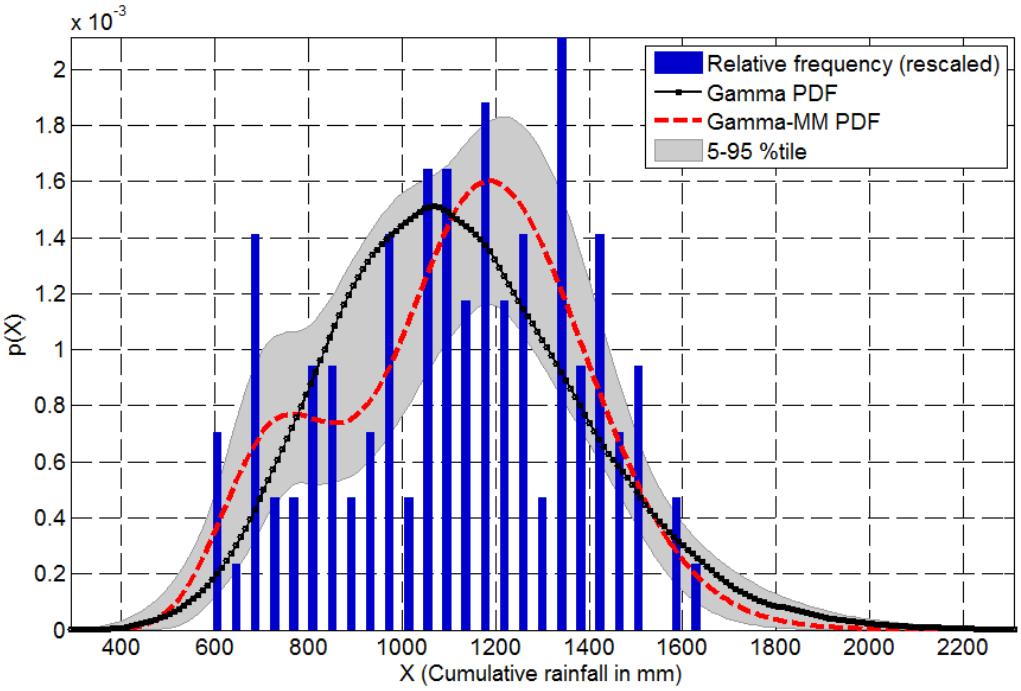

Figure 3.8. Relative frequency of the cumulative rainfall amounts during the south-west summer monsoon months (JJAS) at Grid 125, and probability density functions of the fitted gamma distribution (Gamma PDF) and gamma mixture model (Gamma-MM PDF). The grey band shows 90% credible interval (5^{th} and 95^{th} percentile) of the Gamma-MM PDF.

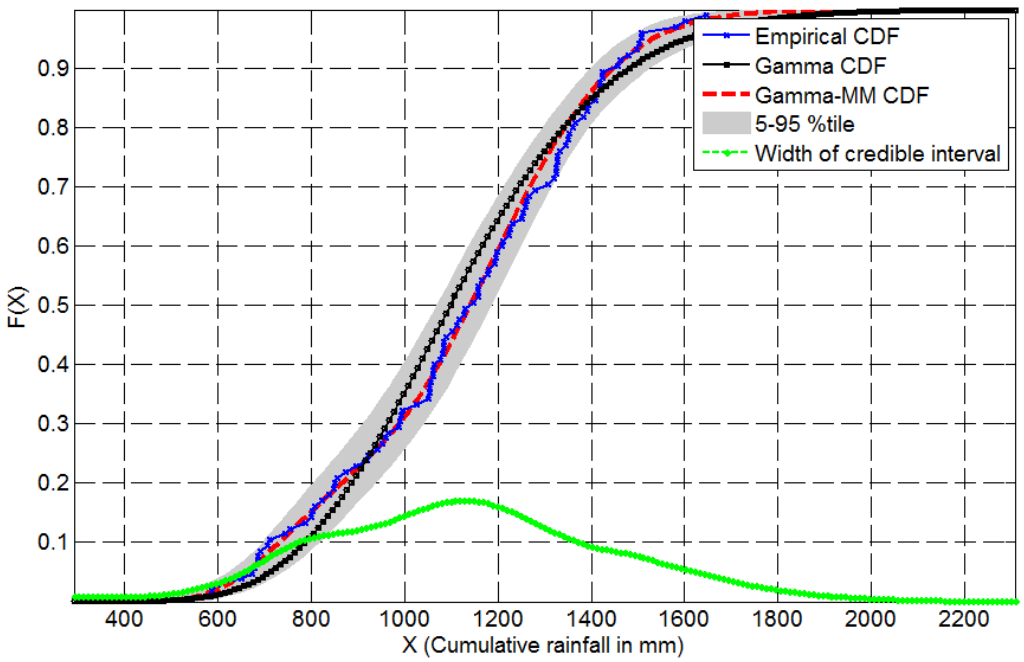

Figure 3.9. Empirical CDF along with CDFs obtained by fitting gamma distribution (Gamma CDF) and gamma mixture model (Gamma-MM CDF) to the cumulative rainfall during the south-west summer monsoon months (JJAS) at Grid 125. The grey band shows 5^{th} and 95^{th} percentile of the Gamma-MM CDF and the green dotted line shows width of its credible interval.

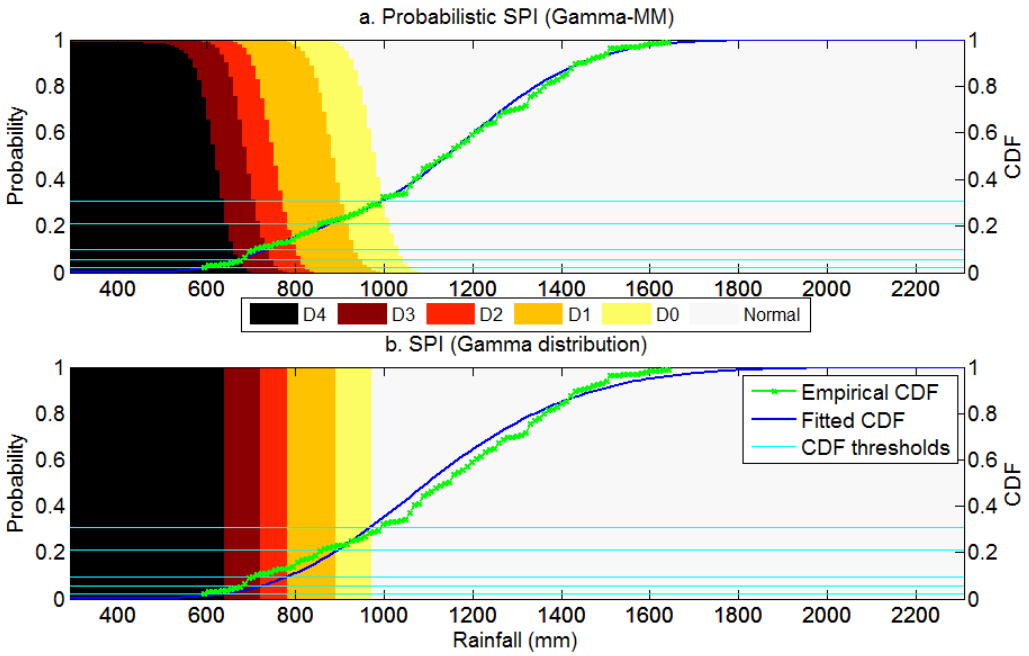

Figure 3.10. Drought classification using rainfall during the south-west summer monsoon months (JJAS) at Grid 125 by the probabilistic SPI (top panel) and standard SPI (bottom panel). The colored patches represent drought classes, the light horizontal lines denote thresholds on CDF specified by US Drought Monitor, and the solid curves represent empirical and fitted CDFs.

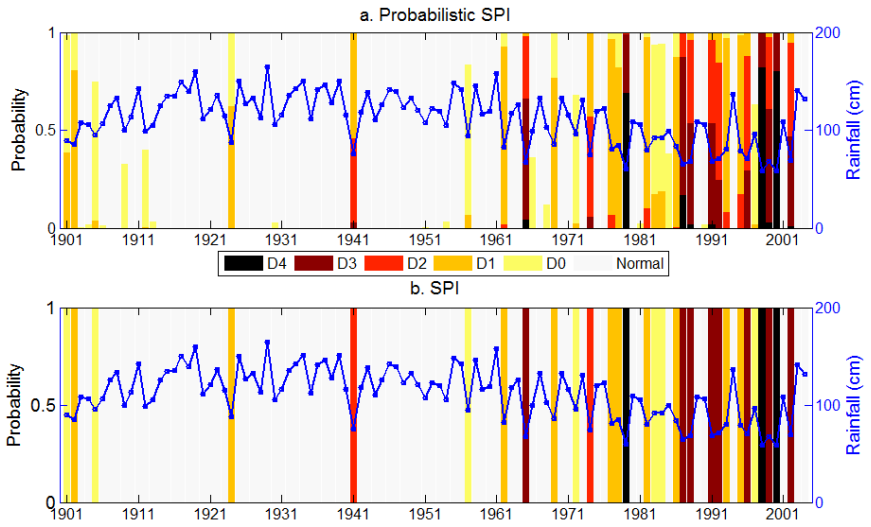

Figure 3.11. Classification of historical droughts during the south-west summer monsoon months (JJAS) at Grid 125 using probabilistic and standard SPI approaches. The solid blue line represents cumulative rainfall during a water-year, a colored bar denotes drought classes and its length represents probability of drought state.

3.5.2 Grid 251 (26°30′ N and 95°30′ E):

This grid point is located in North-East India, which is among the highest rainfall receiving regions of the world. Figure 3.12 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. Figure 3.13 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 (Figure 3.12). Figure 3.14 presents the demarcating boundaries for the drought classes determined by the two methods. A notable feature in the figure is a relatively diffuse 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 mm to 3500 mm, Figure 3.13). The drought classification of the historical data is given in Figure 3.15. Compared to standard SPI, the probabilistic SPI is more conservative in assigning D4 category drought. For example, 1953, 1954 and 1955 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.

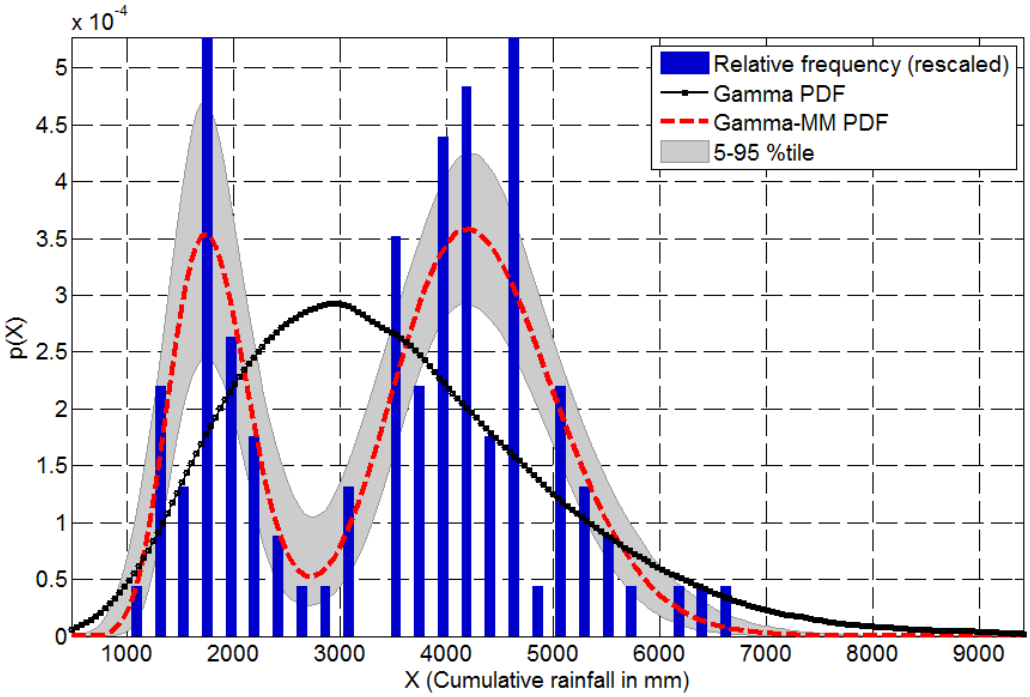

Figure 3.12. Relative frequency of the cumulative rainfall amounts in a water-year at Grid 251 in NE India, and probability density functions of the fitted gamma distribution (Gamma PDF) and gamma mixture model (Gamma-MM PDF). The grey band shows 90% credible interval (5^{th} and 95^{th} percentile) of the Gamma-MM PDF.

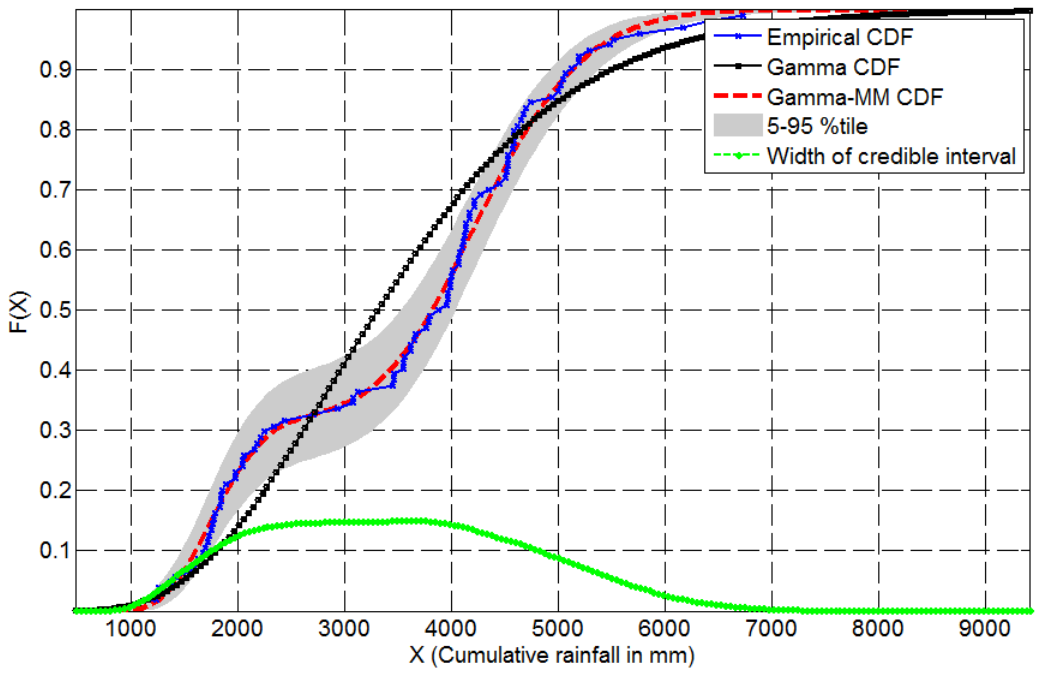

Figure 3.13. Empirical CDF along with CDFs obtained by fitting gamma distribution (Gamma CDF) and gamma mixture model (Gamma-MM CDF) to the cumulative rainfall in a water-year at Grid 251 located in NE India. The grey band shows 5^{th} and 95^{th} percentile of the Gamma-MM CDF and the green dotted line shows width of its credible interval.

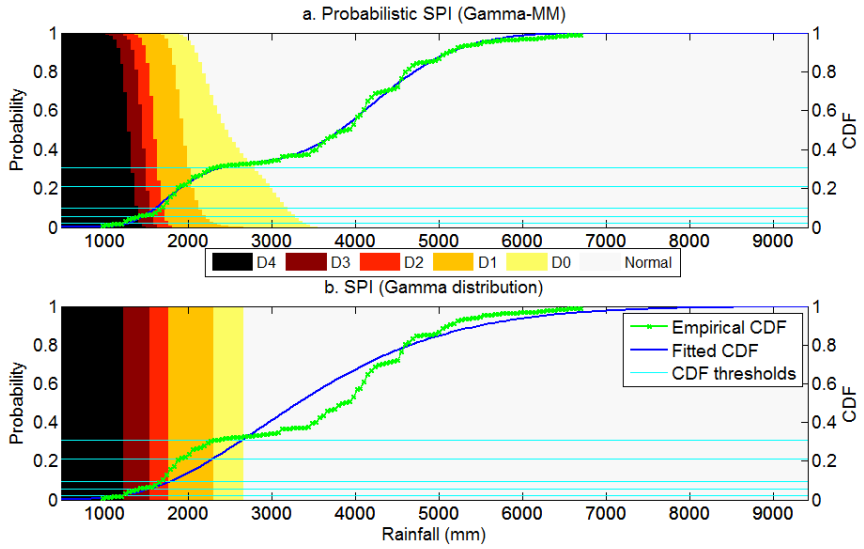

Figure 3.14. Drought classification using rainfall at Grid 251 in NE India by the probabilistic SPI (top panel) and standard SPI (bottom panel). The colored patches represent drought classes, the light horizontal lines denote thresholds on CDF specified by US Drought Monitor, and the solid curves represent empirical and fitted CDFs.

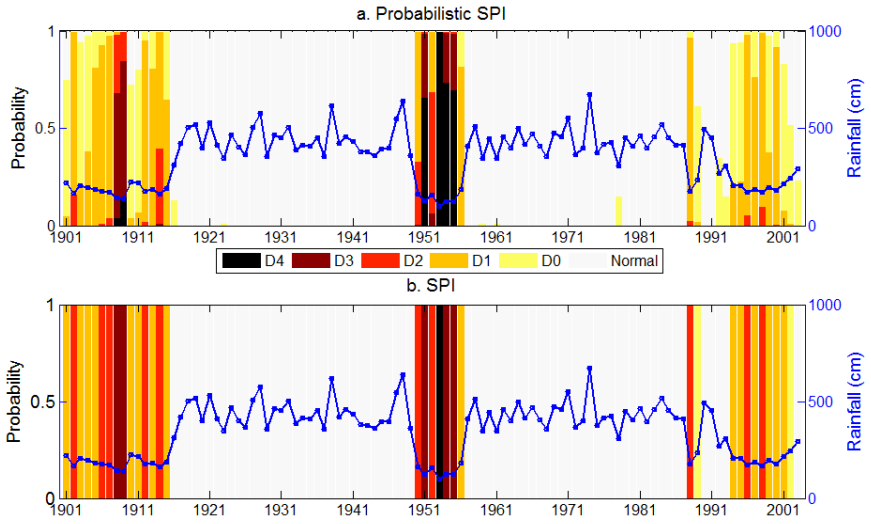

Figure 3.15. Classification of historical droughts during a water-year at Grid 251 in NE India using probabilistic and standard SPI approaches. The solid blue line represents cumulative rainfall during a water-year, a colored bar denotes drought classes and its length represents probability of drought state.

3.5.3 Grid 278 (28°30′ N and 70°30′ E):

This grid point belongs to the Thar Desert in western India where the annual rainfall is much smaller than rest of the country. Figure 3.16 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 (Figures 3.16 and 3.17). 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. Figure 3.18 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 a 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 very similar (Figure 3.19). 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.

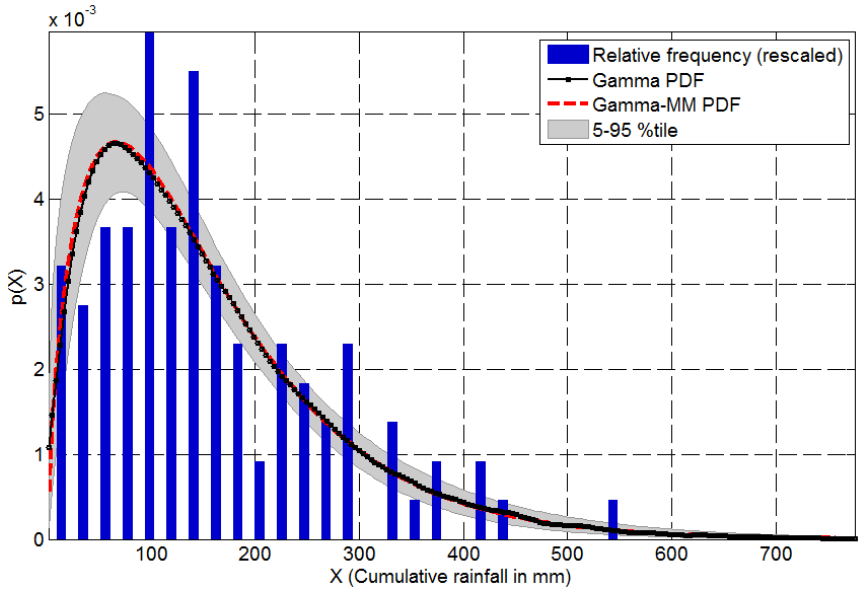

Figure 3.16. Same as Figure 3.12 but for Grid 278 in the Thar Desert of Western India.

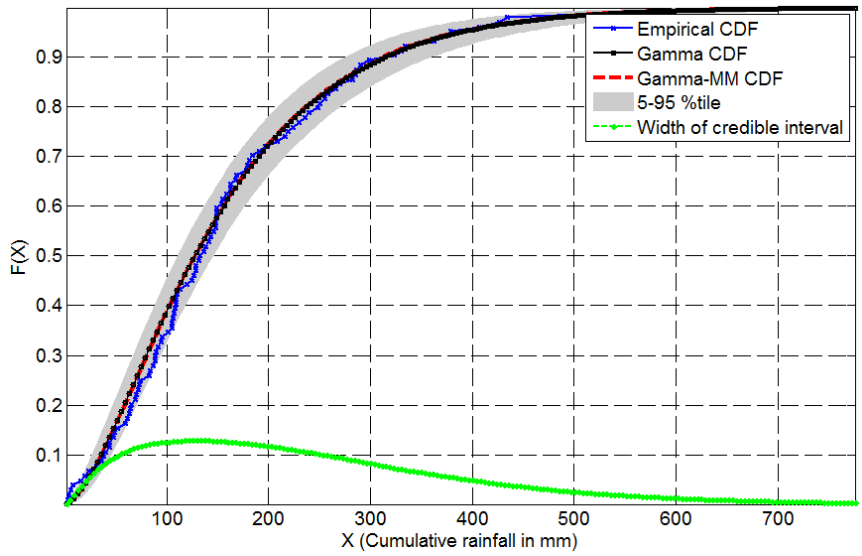

Figure 3.17. Same as Figure 3.13 but for Grid 278 in the Thar Desert of Western India.

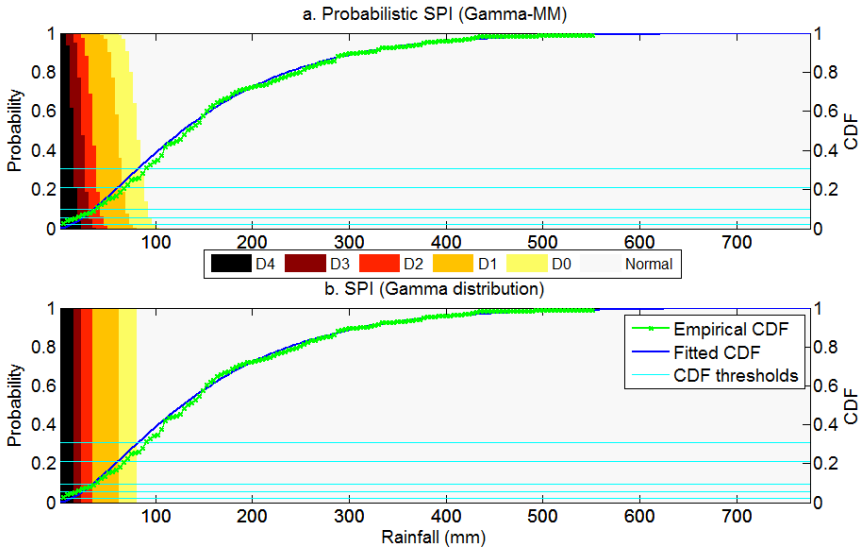

Figure 3.18. Same as Figure 3.14 but for Grid 278 in the Thar Desert of Western India.

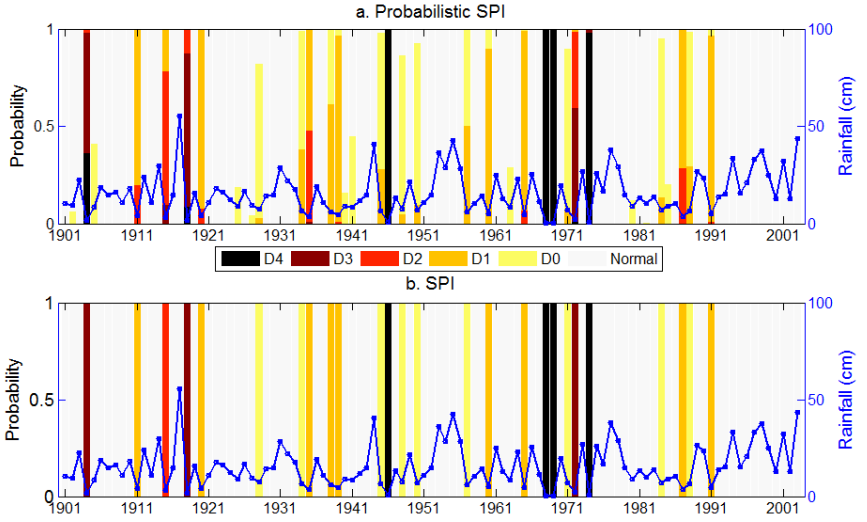

Figure 3.19. Same as Figure 3.15 but for Grid 278 in the Thar Desert of Western India.

3.6 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.

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 the data satisfy SPI assumptions. However, the results of the new method are markedly different and more intuitive than SPI results for situations where the data violate SPI assumptions. The drought classifications obtained using the proposed method were less conservative compared to the probabilistic classifications by HMM with 11 hidden states as the proposed method 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; 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 the HMM-DI, can be a viable tool for these analyses.

In this chapter, a probabilistic SPI was applied to rainfall data. However, the proposed method can be easily extended to classifying droughts using other hydrometeorological variables such as streamflow, runoff, groundwater, and soil moisture for which SPI-like indices have been proposed in the literature. Many of these hydrometeorological variables have large measurement uncertainties, which are ignored in standard SPI type analysis, but can be easily engaged in the proposed method. Further, probabilistic drought classification was being carried out using previously observed precipitation data. However, the Gamma mixture model with known parameters can be used to model future precipitation sequences. The uncertainties associated in the CDF of these precipitation sequences would be referred to as prediction intervals. Note that the prediction intervals will be wider than credible intervals because they account for both the model uncertainty and the noise in precipitation data. Using the framework presented in this chapter, the prediction intervals can then be engaged to obtain probabilistic drought classification for the selected future time period of analysis.

4. IDENTIFICATION OF HOMOGENEOUS DROUGHT REGIONS

This article will be submitted to a suitable journal for publication.

4.1 Introduction

Droughts are complex natural phenomena that are primarily caused due to deficit in precipitation over a region. Due to the ever-increasing water demand for human consumption, agricultural activities, and industries, a lack of water availability during droughts can have devastating effects on all living beings, deteriorate the water quality, as well as hurt the economy (Wilhite, 2000). Also, regions that are most vulnerable to droughts are subject to continual environmental degradation and, when neglected, can lead to desertification (Glantz and Orlovsky, 1983).

Historic droughts are ranked and classified based on drought characteristics such as severity, timing, duration, and spatial extent. Whenever droughts manifest, they last for an extended period and affect large geographic regions, as opposed to the relatively local nature of hydrologic processes such as precipitation and streamflow. Thus the areal extent of droughts is an essential factor for drought management purposes.

It is also not uncommon for a region to be arid while a neighboring region experiences above-normal precipitation. Such a scenario is frequently a result of the presence of geographic features such as mountains, complex atmospheric circulation patterns, and local hydrological processes over the affected area (Bravar and Kavvas, 1991). Therefore, it is a challenge to find areas with similar drought characteristics, especially over a country as large and geographically and climatologically diverse as India. According to Dracup et al. (1980), any process of finding homogeneous drought regions should consider the homogeneity of both climate and geomorphology. Once

contiguous areas having similar drought characteristics are identified, they provide means to pool the hydrometeorological data over the area (Dracup et al., 1980) and analyze the spatial and temporal properties of droughts on a regional scale, leading to efficient drought management (Wilhite and Svoboda, 2000).

There are several regionalization studies for hydrometeorological variables such as precipitation, temperature, etc. For example, Gadgil and Joshi (1983) used principal component analysis on monthly precipitation and temperature series over several stations in India. Puvaneswaran (1990) used factor and cluster analysis on monthly and seasonal meteorological series to find homogeneous regions over Queensland, Australia. Stooksbury and Michaels (1991) used average linkage clustering on daily temperature and precipitation series and other statistics to find homogeneous clusters over the southeastern United States. Leber et al. (1995) used factor analysis and Ward's clustering on monthly and seasonal meteorological data over Tibet. Wards clustering and k-means on several monthly precipitation and temperature indices were used over the northeastern United States by Degaetano (1996). Ahmed (1997) used factor analysis, and Alijani et al. (2008) used Ward's linkage clustering for regionalization of meteorologic data over Saudi Arabia and Iran, respectively. Sahin and Cigizoglu (2012) and Iyigun et al. (2013) used monthly precipitation, temperature, relative humidity, and climate indices to find homogeneous regions over Turkey using neuro fuzzy-based clustering and Ward's clustering. Bharath and Srinivas (2015) used wavelet-based fuzzy c-means cluster analysis for the delineation of similar hydrometeorological regions over India and also highlighted the need to have long records when using principal components as predictors within any clustering algorithm. Their study used monthly time series of precipitation, minimum temperature, and maximum temperature at 0.5° spatial resolution to identify 29 homogeneous hydrometeorological regions. Over each hydrometeorological region, regional drought analysis was performed using standardized precipitation evapotranspiration index (SPEI). Principal component analysis on hydrometeorological variables was used by Aldrian and Dwi Susanto (2003) and Abatzoglou et al. (2009) for regionalization in Indonesia and California,

respectively, and in combination with an artificial neural network by Malmgren and Winter (1999) over Puerto Rico. Clustering is also used in other hydro-meteorological applications such as optimization of networks of weather stations (DeGaetano, 2001), identification of homogeneous regions for regional frequency analysis of precipitation extremes (Kyselỳ et al., 2007), regional flood frequency analysis (Srinivas et al., 2008) and locating climatologically homogeneous regions (Matulla et al., 2003; Unal et al., 2003).

Several studies have used clustering methods to obtain homogeneous drought regions (Stahl and Demuth, 1999; Alvarez and Estrela, 2003; Lana et al., 2001). Vicente-Serrano (2006a) analyzed spatial patterns of droughts using principal component analysis and found that these patterns changed with the time-scale of analysis and reported finding incoherent clusters for 24- and 36-month time-scales. Vicente-Serrano (2006b) used a hierarchical clustering algorithm to find homogeneous drought regions over the Iberian Peninsula. Zhang (2004) used the fuzzy clustering technique to sub-divide maize growing areas based on their vulnerability to droughts. Trnka et al. (2009) used hierarchical clustering to group stations with similar drought characteristics. The usefulness of identifying homogeneous regions was demonstrated in the study by Hisdal and Tallaksen (2003), where severity-area-frequency curves were developed, that allowed the estimation of the return period of drought events. Hannaford et al. (2011) used spatial coherence of droughts between homogeneous regions within Europe to facilitate an early warning system of droughts in a target region, using information about droughts developing elsewhere. Some studies have also looked at the identification of homogeneous drought regions over parts of India. For example, Ghosh and Srinivasan (2016) analyzed spatio-temporal characteristics of droughts over southern peninsular India using k-means, while Goyal and Sharma (2016) used fuzzy c-means approach for finding homogeneous meteorological drought regions over western India. The drought clusters identified in these studies provide useful insights into homogeneous clusters over sub-regions in India.

A challenge encountered in clustering studies is that every algorithm has its strengths and weaknesses (Duda et al., 2012; Fred and Jain, 2005; Kuncheva and Hadjitodorov, 2004). Depending on the choice of the clustering algorithm, the dimensionality of the input dataset and model parameters, cluster assignments can differ. Several cluster ensemble techniques are proposed in the literature where outputs from multiple clustering algorithms, or various parameter settings (e.g., number of clusters K in k-means) of the same algorithm, are combined to obtain final cluster assignments. Among them are (a) *feature-based* methods where clustering results obtained from several base clustering algorithms are used as inputs to another model, thereby turning it to a problem of clustering of cluster labels (Nguyen and Caruana, 2007) (b) *graph partition* methods (Strehl and Ghosh, 2002), and (c) *pairwise similarity* approach (Fred and Jain, 2005; Monti et al., 2003). Studies have shown that irrespective of the approaches listed above, the final cluster assignments were robust when diverse base clustering models were used within the ensemble (Law et al., 2004; Kuncheva and Vetrov, 2006; Monti et al., 2003; Ayad and Kamel, 2003). The link-based similarity method harnesses the simplicity offered by the pairwise similarity method of utilizing similarity between data points, but also extracting the underlying information due to the association between different clusters in the ensemble. The clustering of clusters has been applied in several domains, such as gene expression analysis (Monti et al., 2003), satellite image analysis (Kyrgyzov et al., 2007), social network analysis (Klink et al., 2006), but not in hydrologic or meteorologic studies.

In this study, homogeneous drought regions over India are identified using multiple drought indices and clustering algorithms at different times scales. Two methods of combining the results from different clustering algorithms are presented. Drought characteristics obtained from standardized precipitation index (SPI), standardized precipitation evapotranspiration index (SPEI), and probabilistic SPI (pSPI) are used in this study. Unsupervised clustering methods, namely Markov random field-based graph cuts algorithm, k-means, and hierarchical agglomerative clustering, are used to obtain an initial set of homogeneous drought regions. The homogeneous drought re-

gions from each clustering method are compared in terms of their number and spatial extent. The stability of these clusters for different combinations of inputs (drought characteristics, land-use, geographic information, etc.) and time-scale of analysis (4-month, 12-month, etc.) is examined. The evolution of homogeneous drought regions over time is studied by performing clustering for different epochs. Finally, the clustering of homogeneous drought clusters obtained from different base clustering methods is presented using (i) similarity matrix and (ii) connected triple-based similarity approaches. Regional drought characteristics are studied by appropriately pooling the data from resulting homogeneous drought clusters, and products such as intensity-duration-frequency curves and intensity-area-frequency curves are developed for each region. The results reported in this study are expected to help policy-makers and water resources managers in developing effective water management and drought mitigation plan for each region.

4.2 Study area and data used

The geographical region chosen for this study is India (Figure 4.1). The study area comprises of the following broad climate regimes: arid (northwest India), semi-arid (Indo-Gangetic plains, central, and interior peninsular India), humid (coastal areas, southern peninsula and northeastern India) and alpine (Himalayan ranges in the north) regions. Although in varying amounts, the majority of India receives precipitation during the southwest monsoon (June to September), while some of the southeastern states receive precipitation during the northeast monsoon season (October to December). When considering temperature across India, winter months typically span between December to February, with northern India experiencing lower temperatures compared to southern India. Summer months typically span between March to May, with the highest temperatures recorded in northern and western parts of India. Based on monsoon precipitation data collected over 306 representative stations, the Indian Institute of Tropical Meteorology (IITM) divides the study area into

five homogeneous monsoon regions (Mooley et al., 1981; Kothawale and Rajeevan, 2017). Figure 4.2 shows the average annual precipitation totals (June to May) for the water years 1902 to 2004 over the study area. The precipitation total during the water year 1902 is computed by summing monthly precipitation total between June 1901 to May 1902, and so on. The mean, maximum, minimum, and standard deviation of these precipitation totals are 1085.5 mm, 1333.6 mm (during 1918 water year), 831.4 mm (during 2003 water year), 101.8 mm, respectively. The histogram of water year precipitation totals for the five homogeneous monsoon regions of the entire study area are shown in Figure 4.3. During the period 1901-2004, northwestern India receives the least precipitation with a mean annual total of 539.5 mm and a standard deviation of 142 mm (Figure 4.3a). Similarly, northeastern India receives the highest precipitation with a mean annual precipitation total of 2068 mm and a standard deviation of 174 mm (Figure 4.3b).

Gridded daily precipitation data from the India Meteorological Department (IMD) (Rajeevan, 2006) available over 355 grids for the period 1901 to 2004 at $1^\circ \times 1^\circ$ spatial resolution (Figure 4.1) were used in this study. The daily precipitation series at each grid were aggregated to obtain monthly precipitation data. A higher resolution precipitation dataset from IMD (Pai et al., 2014) available at 4913 grids over India at a resolution of $0.25^\circ \times 0.25^\circ$ from 1901 to 2019 was used to test the sensitivity of regionalization results to the spatial resolution of the inputs provided to clustering algorithms. Drought characteristics at multiple time scales (4-month and 12-month time windows) using SPI (McKee et al., 1993), SPEI (Vicente-Serrano et al., 2010), and pSPI (Mallya et al., 2013) were computed. Other datasets used for this study include: $0.5^\circ \times 0.5^\circ$ gridded temperature data for SPEI analysis (UDelAirTPrecip, http://www.esrl.noaa.gov/psd/), and land cover data was obtained from Roy et al. (2016) (https://tinyurl.com/vp6txe5). The land use and land cover data (LULC) over India with a spatial resolution of 100m is shown in Figure 4.4 and corresponds to the year 2005. The data corresponding to only the year 2005 was used in the study. According to this LULC dataset, the following are the top 5 land use classes over the

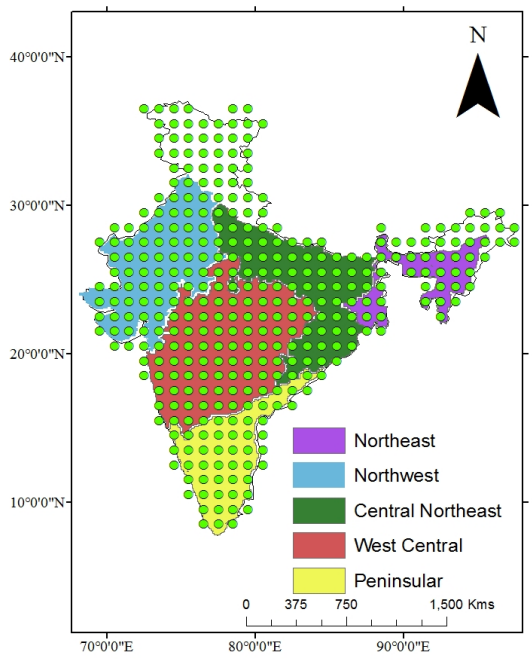

Figure 4.1. Study area with 1° × 1° India Meteorological Department (IMD) precipitation grids shown as green circular markers. Homogeneous monsoon regions (IITM) over India are shown in the background.

study area: Corp land (48%), Deciduous broadleaf forests (9%), Fallow land (14%), Evergreen broadleaf forests (12%), and Shrubland (11%). A detailed legend for LULC classes shown in Figure 4.4 is given in Table 4.1. The percentage of each land use class within 1° × 1° grid over India was computed using geographic information software (GIS).

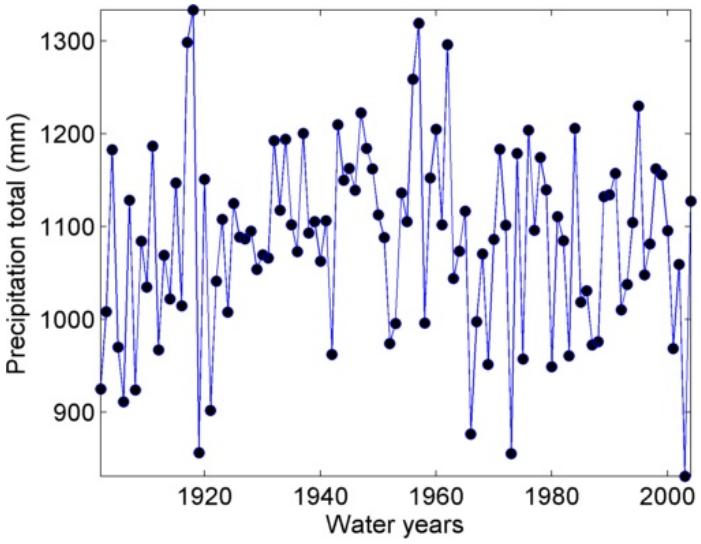

Figure 4.2. Average water year precipitation time series over India.

Table 4.1.
Land Use and Land Cover classes across India

No.	**Description**	No.	**Description**
1	Deciduous Broadleaf Forest	11	Aquaculture
2	Crop land	12	Mangrove Forest
3	Built-up Land	13	Salt Pan
4	Mixed Forests	14	Grassland
5	Shrubland	15	Evergreen Broadleaf Forest
6	Barren land	16	Deciduous Needleleaf Forest
7	Fallow land	17	Permanent wetland
8	Wasteland	18	Snow and Ice
9	Water bodies	19	Evergreen Needle forest
10	Plantations		

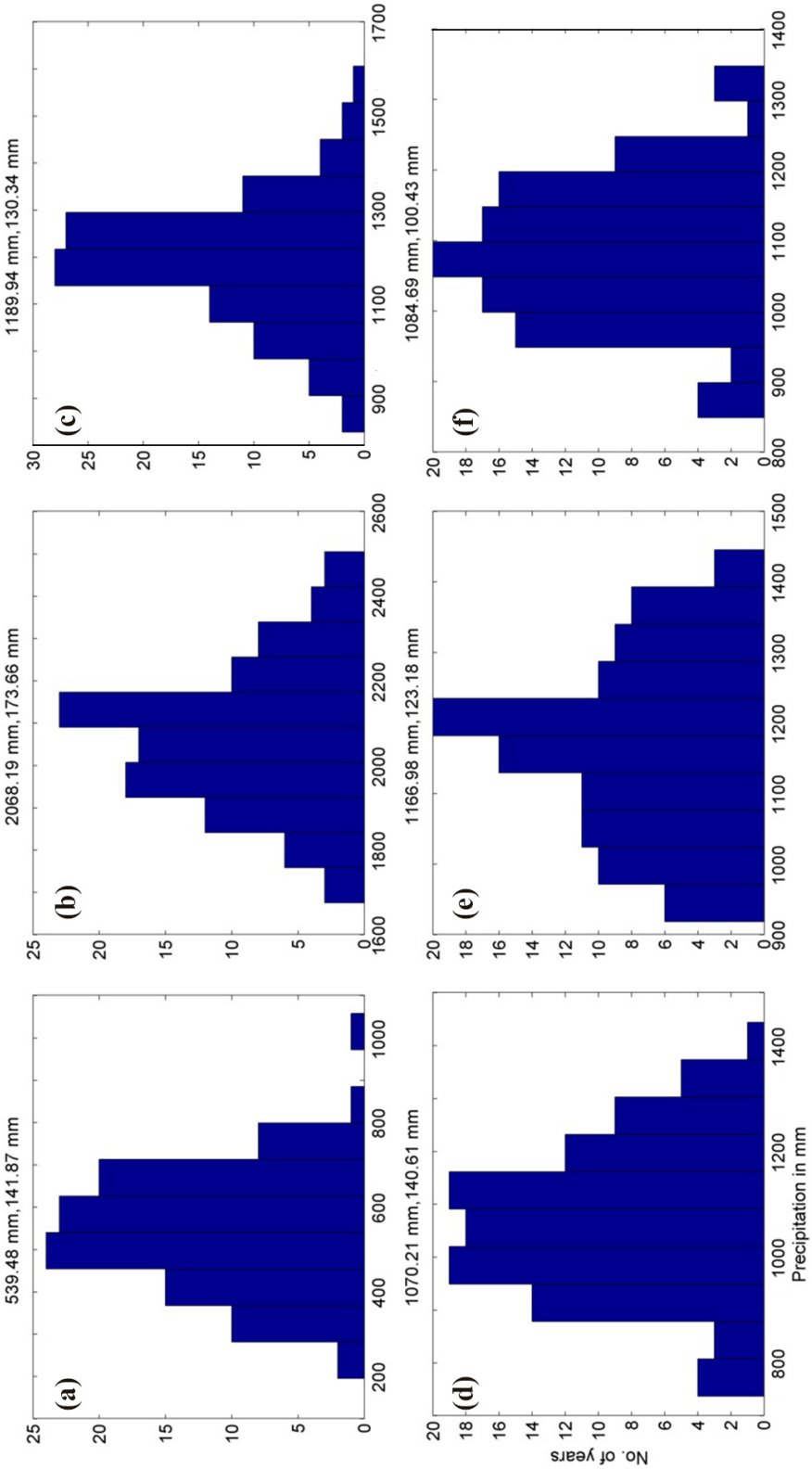

Figure 4.3. Histogram of annual precipitation totals during the period 1901-2004 over the five Homogeneous monsoon regions (IITM) (a) Northwest (NW), (b) Northeast (NE), (c) Central Northeast (CNE), (d) West Central (WC), and (e) Pensinsular India (PI). The mean and standard deviation are noted in the title of each subplot.

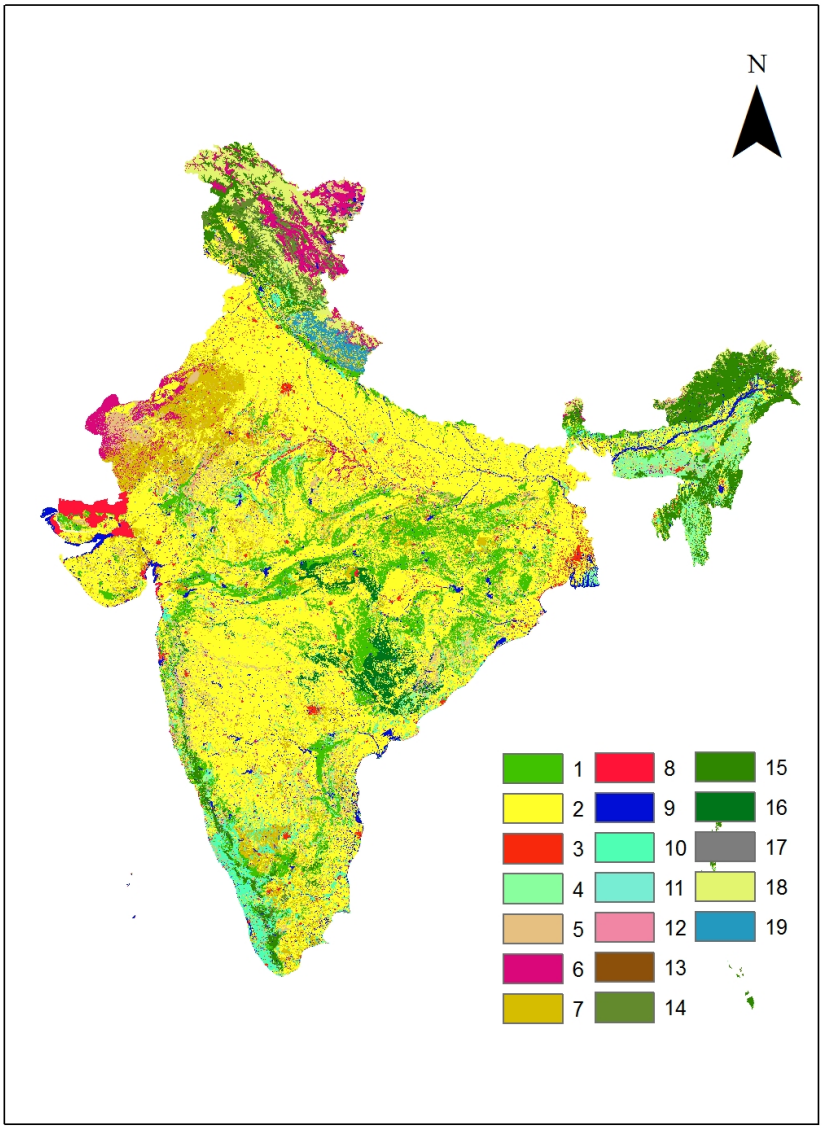

Figure 4.4. Land Use and Land Cover classification over India during 2005.

4.3 Methodology

4.3.1 Clustering using k-means

The k-means is a clustering algorithm that groups data available at G grids or stations over the study area into K groups of equal variance by minimizing the within-cluster sum of squared deviations about the centroid. The number of clusters, K, for k-means algorithm has to be pre-specified. The objective function may be written as:

$$\underset{\mathbf{S}}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{\mathbf{x} \in S_i} ||\mathbf{x} - \mu_i||^2 \tag{4.1}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_G$ are observations over G grids, and each observation can be a d-dimensional real vector. The observations $\mathbf{x}_1, \ldots, \mathbf{x}_G$ are divided into K sets $\mathbf{S} = \{S_1, \ldots, S_K\}$, with μ_i representing the mean of data belonging to set S_i.

The k-means algorithm works as follows:

1. Pre-specify the number of clusters K in the model.
2. Randomly select K data points as the centroids.
3. Calculate the distance of each data point $\boldsymbol{x}$ to each of the K centroids, and assign the data point to the cluster whose centroid is the closest.
4. Calculate the position of the centroid based on the cluster assignments obtained in step 3.
5. Repeat steps 3 and 4 until there is no change in cluster assignment or maximum iterations has reached. Typical value of maximum iterations is 100.

In many real-world applications, like the one discussed in this chapter, the number of clusters is not known before hand. Therefore, k-means clustering is repeated for several values of K, and in each case model performance statistics such as Davis-Bouldin (DB) index, Silhouette coefficient, and Bayesian information criteria (BIC) score are computed.

The Davies-Bouldin (DB) index is defined as follows (Davies and Bouldin, 1979):

$$DB = \frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j} R_{ij} \tag{4.2}$$

where R_{ij} is a similarity metric defined such that it provides trade off between (a) the average distance (s_i) between each point of cluster i and its centroid (i.e. scatter within cluster i), and (b) the distance (d_{ij}) between cluster centroids i and j. This similarity metric is calculated as follows:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{4.3}$$

The Silhouette coefficient (Rousseeuw, 1987) provides a measure of how similar a data point is to other members of its own clusters compared to members of other neighboring clusters and is calculated as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |K_i| > 1 \tag{4.4}$$

where $s(i)$ is the silhouette value of data point i belonging to cluster K_i and $|K_i|$ is the number of data points assigned to that cluster. Also $a(i)$ is the mean distance of the data point i to all other data in K_i and is given as:

$$a(i) = \frac{1}{|K_i| - 1}\sum_{j\in K_i, i\neq j} d(i,j) \tag{4.5}$$

where $d(i,j)$ is the distance between data points i and j within cluster K_i. Similarly, $b(i)$ in equation 4.4 is the smallest distance between data point i and all points in any other cluster to which i is not assigned.

$$b(i) = \min_{m\neq i}\frac{1}{|K_m|}\sum_{j\in K_m} d(i,j) \tag{4.6}$$

The BIC (Schwarz, 1978) is defined as follows:

$$BIC = G\ln\frac{RSS}{G} + K\ln(G) \tag{4.7}$$

where G is the total number of grids used for regionalization, K is the number of homogeneous regions, and RSS is the residual sum of squares of the regionalization model with K regions. The best model is the one with the smallest BIC value.

Therefore, the k-means model with the least DB index or the least BIC value or the highest Silhoutte coefficient is chosen as the optimal model, and the optimal K is noted.

4.3.2 Clustering using Markov Random Fields

The aim of regionalization is to find groupings (or clusters) of grid points such that drought characteristics of grid points within a group are similar. To achieve this aim, a metric is required that can quantify the similarity between the grid points based on drought characteristics. Spatial contiguity of the clusters is also desired in regionalization as droughts tend to be continuous in space. These two factors, similarity metric and spatial contiguity, are addressed through a regionalization algorithm developed in a Bayesian framework by using the concepts of Gaussian mixture model (GMM) and Markov random fields (MRF), and is described below.

Let the drought state at grid l be given by $\mathbf{Z}_l, l = 1, \ldots, G$ ($\mathbf{Z}_l = \{z_1, \ldots, z_M\}$ and $Z = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_G\}$) where G is the total number of grid points and M is the dimensionality of the input dataset. Let C be the number of clusters or regions, and $\mathbf{f} = [f_1, \ldots, f_G]^T, f_l \in 1, \ldots, C$ be the cluster labels. The posterior distribution of the cluster labels is obtained by using Bayes' rule as:

$$P(\mathbf{f}|Z) = \frac{P(Z|\mathbf{f})P(\mathbf{f})}{P(Z)} \tag{4.8}$$

where $P(\mathbf{f})$ is the prior probability of the cluster labels and $P(Z|\mathbf{f})$ is the likelihood function estimated as given by Tripathi and Govindaraju (2009):

$$P(Z|\mathbf{f}) = \prod_{l=1}^{G} P(Z_l|f_l) = \prod_{l=1}^{G}\prod_{c=1}^{C} N(\mathbf{Z}_l|\boldsymbol{\mu}_{Z_c}, \boldsymbol{\Sigma}_{Z_c})^{I_{lc}} \tag{4.9}$$

In Equation 4.9 I_{lc} is an indicator variable taking a value of one if $f_l = c$, and zero otherwise, and $\boldsymbol{\mu}_{Z_c}$ and $\boldsymbol{\Sigma}_{Z_c}$ are the mean and covariance of the drought (or wet) state in cluster c. The probability $N(\mathbf{Z}_l|\boldsymbol{\mu}_{Z_c}, \boldsymbol{\Sigma}_{Z_c})$ will be large if the duration, magnitude, and timing of the drought states in grid l are similar to those of the region c, and small otherwise, thus providing the desired similarity metric.

To encode the preference for spatial contiguity in the clusters, the prior probability $P(\mathbf{f})$ in Equation 4.8 is chosen to be a Markov random field (MRF). A MRF, also known as a Markov network, is commonly used to model the joint distribution between spatially dependent variables. Using the Hammersley-Clifford theorem (Clifford, 1990), $P(\mathbf{f})$ is expressed as:

$$P(\mathbf{f}) = \frac{1}{\Xi} exp\left[-\sum_{l,m \in N} E\left(f_l, f_m\right)\right] \tag{4.10}$$

where Ξ is a normalizing constant, E is an energy function in the space of clusters, and N is a neighborhood set for the grid points. The prior distribution encodes the belief that the grid points in the neighborhood set are more likely to have same cluster labels - the degree of belief being controlled by the energy function. The energy function was selected as:

$$E\left(f_l, f_m\right) = \begin{cases} 0, & \text{if } f_l = f_m, \forall\ l, m \in N \\ \alpha, & \text{if } f_l \neq f_m, \forall\ l, m \in N \end{cases} \tag{4.11}$$

A value of $\alpha = 0$ corresponds to a uniform distribution in the space of clusters, i.e. no preference for spatial contiguity is implied and $P(\mathbf{f})$ is non-informative. Higher values of α forces neighbors to be in the same clusters. It was observed that the choice of prior distribution, $P(\mathbf{f})$, essentially affects the cluster assignment of primarily those grid points that are either on the fringes of a region or are outliers.

The posterior probability with an MRF prior cannot be estimated analytically. In this study, for simplicity, the posterior distribution was approximated by a delta function at its mode $P(\mathbf{f}|Z) : \delta\left(\mathbf{f}^{\hat{a}}\right)$ by maximizing the logarithm of the posterior distribution:

$$\mathbf{f}^{\dot{a}} = \underset{f}{\operatorname{argmax}} \sum_{l=1}^{G} \sum_{c=1}^{C} \ln P(\mathbf{Z}_l | f_l, \boldsymbol{\mu}_{Z_c}, \boldsymbol{\Sigma}_{Z_c}) - \sum_{l,m \in N} E\left(f_l, f_m\right) \tag{4.12}$$

The maximization is performed iteratively following the graph cuts algorithm (Boykov et al., 2001) that guarantees the solution to lie within a constant factor from the global maximum. The steps involved in maximizing the objective function are described in detail in (Tripathi and Govindaraju, 2009). As in the case of k-means, the clusters formed are evaluated using Davies-Bouldin (DB) index or Silhouette coefficient or Bayesian information criterion (BIC).

4.3.3 Hierarchical clustering

Hierarchical clustering builds nested clusters by merging and splitting them iteratively until some prespecified criterion is satisfied. The hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree denotes a single cluster with all samples. The leaves of the tree represent clusters with a single sample. Agglomerative clustering (Gowda and Krishna, 1978) is a type of hierarchical clustering where each sample initially represents a cluster. The clusters are then merged together according to the linkage criteria. Different linkage criteria may be adopted, such as:

1. Ward: minimizes the sum of squared differences within clusters, thus minimizing the variance. This objective function is similar to k-means but clustering is performed in a hierarchical manner.

2. Maximum linkage: minimizes the maximum distance between observations of pairs of clusters.

3. Average linkage: minimizes average of distance of all observations between pairs of clusters.

4.3.4 Pairwise similarity matrix

Each of the three methods described above is likely to provide a unique set of clusters, leading to the next challenge of combining the unique clustering results into a meaningful single output. This problem is referred to as clustering of clusters. To address this problem, a pairwise similarity matrix is constructed for each of the M base clusters following Fred and Jain (2005). Given a dataset $\boldsymbol{x} = x_1, \ldots, x_N$, we obtain M base clustering results $\Pi = \{\pi_1, \ldots, \pi_M\}$. Following this, a $N \times N$ similarity matrix S_m, $m = 1, \ldots, M$ is constructed for each base clustering method π_m. When two data points are assigned to the same cluster, the corresponding entry in the similarity matrix will be equal to one, and zero otherwise.

$$S_m(x_i, x_j) = \begin{cases} 1, & \text{if } C(x_i) = C(x_j) \\ 0, & \text{otherwise.} \end{cases} \tag{4.13}$$

Next, the M similarity matrices are combined to form a consensus matrix (Q) (Monti et al., 2003), such that each element within this matrix represents the degree of similarity between any two data points x_i and $x_j \in \boldsymbol{x}$.

$$Q(x_i, x_j) = \frac{1}{M} \sum_{m=1}^{M} S_m(x_i, x_j) \tag{4.14}$$

Finally, this consensus matrix is used as an input to hierarchical algorithm to obtain the final clustering (Fred and Jain, 2005).

4.3.5 Connected-triple-based similarity matrix

Another approach used for clustering of clusters is the use of connected triple-based similarity (CTS) matrix proposed by Klink et al. (2006) and Iam-on et al. (2010). Using this approach, it is possible to combine clusters obtained from each base clustering algorithm such as k-means, graph cuts, agglomerative clustering, etc. In Figure 4.5, the cluster ensemble Π consists of results from m base clustering methods,

$\pi_m, m = 1, \ldots, 3$ when applied to data points $\boldsymbol{x} = \{x_i, \ldots, x_N\}$, where N is the total number of data points. The square nodes in Figure 4.5 denote the clusters produced by each base clustering method. For example, $\{C_1^1, C_2^1, C_3^1, C_4^1, C_5^1\}$ represents five clusters formed when π_1 is used as the based clustering algorithm to data $\boldsymbol{x}$. There exists an edge between data point x_i and clusterC_j^m if x_i belongs to cluster C_j^m for base clustering π_m. From the illustration, x_1 and x_2 are similar according to base clustering methods π_2 and π_3 because both these data points are assigned to the same clusters (i.e. x_1 and $x_2 \in C_1^2$ and x_1 and $x_2 \in C_1^3$). However, if we only consider the cluster assignments in base clustering method π_1, it appears that data points x_1 and x_2 are dissimilar, as they are assigned to clusters C_1^1 and C_2^1, respectively. However, because C_1^1 and C_2^1 possess two connected triples with C_1^2 and C_1^3 as the triple centres, we can show that there is indeed some similarity between x_1 and x_2.

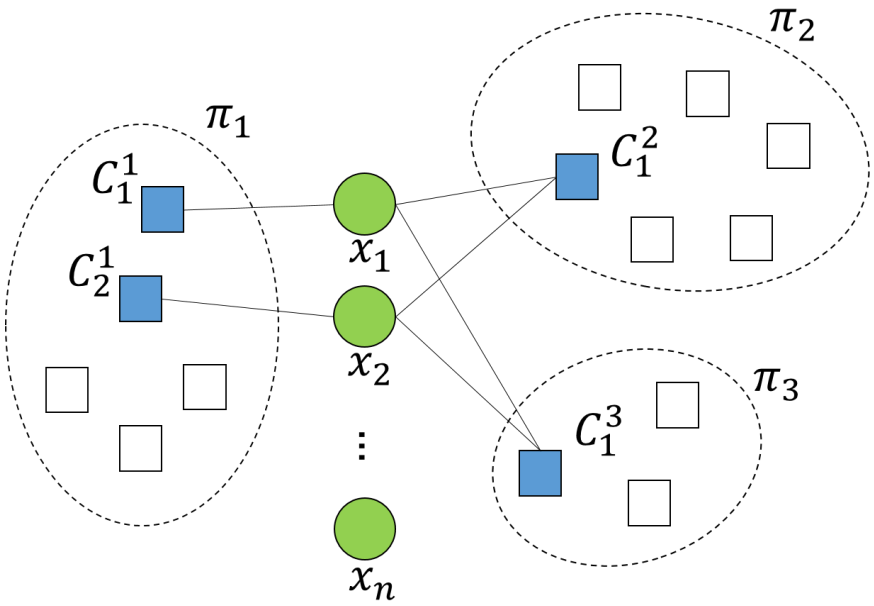

Figure 4.5. A graphical illustration of a cluster ensemble $\Pi = \{\pi_1, \pi_2, \pi_3\}$, where π_1=$\{C_1^1, C_2^1, C_3^1, C_4^1, C_5^1\}$, π_2=$\{C_1^2, C_2^2, C_3^2, C_4^2, C_5^2, C_6^2\}$, and π_3=$\{C_1^3, C_2^3, C_3^3\}$

In this study, the weighted connected triple (WCT) algorithm implemented by Iam-on et al. (2010) was used to compute the similarity matrix. According to WCT algorithm, given a set of data points $\boldsymbol{x}$ and cluster ensemble Π, a weighted graph $G = (V, W)$ can be constructed, where V is a set of vertices representing clusters in Π and W is the weighted edges between clusters. The similarity $Sim^{WCT}(i, j)$ between clusters C_i and C_j is given as:

$$Sim^{WCT}(i,j) = \frac{WCT_{ij}}{WCT_{max}} \tag{4.15}$$

where WCT_{max} is the maximum WCT value between any two clusters in the cluster ensemble Π, and WCT_{ij} is the WCT value between clusters C_i and $C_j \in V$ and is calculated as:

$$WCT_{ij} = \sum_{k=1}^{q} WCT_{ij}^{k} \tag{4.16}$$

where WCT_{ij}^{k} is the number of connected triples between C_i and C_j whose common neighbor is cluster $C_k \in V$ and we count all q $(1 \leq q < \infty)$ triples between C_i and C_j. The number of connected triples WCT_{ij}^{k} can be computed as:

$$WCT_{ij}^{k} = min(w_{ik}, w_{jk}) \tag{4.17}$$

where w_{jk} is the weight of the edge connecting cluster C_j and $C_k \in V$ and is given as of number of data points that are assigned to both clusters to the total number of data points assigned to each cluster.

$$w_{jk} = \frac{|\mathbf{x}_{C_j} \cap \mathbf{x}_{C_k}|}{|\mathbf{x}_{C_j} \cup \mathbf{x}_{C_k}|} \tag{4.18}$$

Therefore, for any ensemble member or base clustering approach $\pi_m \in \Pi$, $m = 1, \ldots, M$, the similarity between data points x_i and $x_j \in \mathbf{x}$ is given by:

$$S_m(x_i, x_j) = \begin{cases} 1, & \text{if } C(x_i) = C(x_j) \\ Sim^{WCT}(C(x_i), C(x_j)) \times DC, & \text{otherwise.} \end{cases} \tag{4.19}$$

where $DC \in (0, 1]$ is a constant decay factor, and denotes the confidence level of acceptance of two non-identical objects as being similar. In this study the decay factor was fixed to 0.6.

4.3.6 Drought Indices

4.3.6.1 Standardized Precipitation Index (SPI)

The SPI is a commonly used drought index to quantify deficit in precipitation (McKee et al., 1993). Depending on the end use, SPI can be calculated for multiple time windows. Typically, SPI at shorter time windows, for example 1-month SPI to 4-month SPI, are useful for identifying and characterizing meteorological and agricultural droughts (Guttman, 1998). Similarly, SPI at longer time windows, for example 6-month SPI, 12-month SPI, etc. are useful for analyzing hydrological droughts (Mishra and Singh, 2010; Mo, 2008). If monthly precipitation time series at a grid is available, in order to compute 4-month SPI a new cumulative precipitation time series is constructed by summing the monthly precipitation amounts for first four months, then summing monthly precipitation for months 2 to 5, and then for months 3 to 6, and so on. Now for each ending month, say 4-month time window ending in September, the 4-month cumulative precipitation time-series ending is September was used to fit a probability distribution function, and then normalized using a standard inverse Gaussian function to obtain SPI values. Moderate, severe, and extreme drought classes were identified using the SPI thresholds specified in Table 4.2. Negative values of SPI indicates drought or deficit in precipitation and positive value indicates above median precipitation or non-drought conditions.

4.3.6.2 Standardized Precipitation Evapotranspiration Index (SPEI)

SPEI (Vicente-Serrano et al., 2010) first requires the computation of potential evapotranspiration (PET). Thornthwaite's equation (Thornthwaite, 1948) was used

Table 4.2.
Drought classification scheme. SPI ranges are prescribed for the inverse of the Normal distribution. Corresponding thresholds on CDF are given in the last column

Description	**SPI Range**	**Threshold on CDF**
Moderate Drought	-1.0 to -1.49	0.07 to 0.16
Severe Drought	-1.5 to -1.99	0.023 to 0.07
Extreme Drought	-2.0 to less	0.023 or less

for computing PET, but other popular approaches may also be used (Penman, 1948; Priestley and Taylor, 1972; Allen et al., 1998). After subtracting the PET from precipitation, SPEI may be computed using similar approach as SPI. (McKee et al., 1993).

4.3.6.3 Probabilistic Standardized Precipitation Index (pSPI)

As discussed in Chapter 3 and Mallya et al. (2015), the problem of choosing an appropriate distribution for SPI analysis can be addressed 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 $(0, \infty)$ (DeVore and Lorentz, 1993). Mallya et al. (2015) provide the mathematical details of probabilistic SPI (pSPI) using Gamma-MM. In brief, if the number of components of the Gamma mixture, M, are known *a priori*, then the weighted sum of M mixtures is given as follows:

$$p(x_t|\lambda) = \sum_{i=1}^{M} w_i G\left(x_t|\nu_i, \frac{\nu_i}{\mu_i}\right) \tag{4.20}$$

where $x_t,\ t = 1, \ldots, N,\ \{x_t \in R$ and $X = [x_1, \ldots, x_N]^T\}$ is the cumulative rainfall at time t, w_i are the mixture weights or mixing ratios, and $G\left(x_t|\nu_i, \frac{\nu_i}{\mu_i}\right)$ are the components of Gamma densities of the form,

$$G\left(x_t|\nu_i, \frac{\nu_i}{\mu_i}\right) = \frac{\left(\frac{\nu_i}{\mu_i}\right)^{\nu_i}}{\Gamma(\nu_i)} x_t^{(\nu_i - 1)} \exp\left(-\frac{\nu_i}{\mu_i} x_t\right) \tag{4.21}$$

with mean μ_i, shape parameter ν_i, and Gamma function $\Gamma(\nu_i)$. Further, the mixture weights satisfy the constraint $\sum_{i=1}^{M} w_i = 1$. The parameter set is represented as, $\lambda = \{w, \mu, \nu\}$, where $w = [w_1, w_2, \ldots, w_M]^T$, $\mu = [\mu_1, \mu_2, \ldots, \mu_M]^T$ and $\nu = [\nu_1, \nu_2, \ldots, \nu_M]^T$. The Gibbs sampler (Geman and Geman, 1984), a Markov-chain Monte Carlo (MCMC) algorithm is used to sample the posterior distribution of the parameters.

4.3.7 Drought characteristics

Drought characteristics are computed on monthly time series of drought intensity values obtained from different drought indices using run theory (Yevjevich, 1967). A *run* is defined as the period of time during which the drought intensity values remain below or above a selected threshold. When the drought intensity values are above the threshold (x_t, say SPI > -1.0), they are denoted as non-drought events and when they are below the threshold they are denoted as drought events (Figure 4.6). A drought event is said to have begun when the drought intensity value falls below the threshold for the first time, and is said to have ended when it goes above the threshold. In Figure 4.6, there are two drought events, d_1 and d_2. Drought duration (D) is defined as the time period over which a drought event persists. For example, in Figure 4.6 the duration of the first drought event (d_1) is D_1. The average drought intensity (I) is computed as the mean of drought intensity values during a drought event. These drought characteristics can be computed for each station (or grid) or for an entire region by pooling data available across stations within the region.

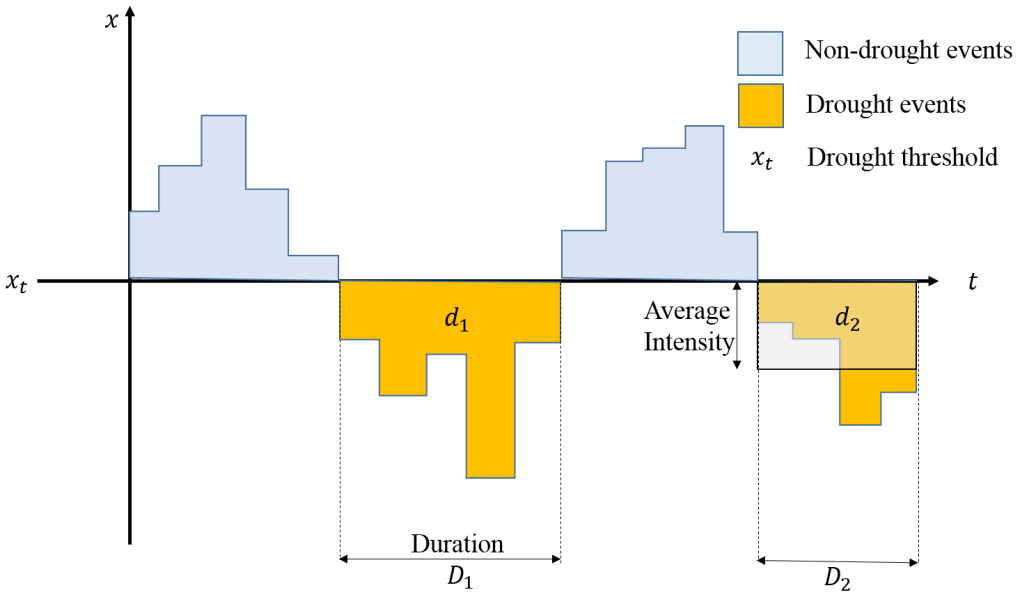

Figure 4.6. Drought characteristics according to run theory.

4.3.8 Regional Intensity-Duration-Frequency analysis

For any selected region, drought events of same duration (for example 3-month long drought event) were identified over all stations within the region. Next, annual minimum series of average drought intensity for the selected duration was determined for the region for frequency analyses. As SPI values during drought events are negative, the minimum value (i.e., highest intensity droughts) is equivalent to maxima values that are used during frequency analyses of precipitation or streamflow. Several candidate distributions such as Gumbel, Generalized Extreme Value (GEV), and t-location-scale were tested for goodness of fit using Chi-square test with a significance level (α) of 1%. The GEV distribution was found to provide the best overall fit. The parameters of the distribution were then used to obtain values of drought intensity for different return periods.

4.3.9 Regional Intensity-Area-Frequency analysis

The spatial and temporal characteristics of droughts over different regions in India were analyzed using drought intensity values recorded at different grids within the region. For a pre-specified drought index and time window (e.g. SPI - 12 month window) the areal extents of drought and spatially averaged drought intensity values were tabulated for all years using GIS over each homogeneous drought region in India (Loukas and Vasiliades, 2004). Frequency analyses was then performed to develop Intensity-Area-Frequency curves for different return periods of interest over the region.

4.4 Results and discussion

The goal of this study was to identify regions over India that have similar drought characteristics. Therefore, drought characteristics over each $1^\circ \times 1^\circ$ grid over India were computed using three drought indices, namely, standardized precipitation index (SPI), standardized precipitation evapotranspiration index (SPEI), and probabilistic SPI (pSPI). Drought characteristics were computed for a 12-month time window ending in May (i.e., water year in India) and 4-month time window ending in September (southwest monsoon season). In addition to using drought characteristics at each grid, their latitude and longitude, as well as land use characteristics, were used as predictors for clustering algorithms.

In this section, homogeneous drought regions obtained from three *base* clustering algorithms are first presented. Each clustering algorithm produces a unique set of clusters due to the inherent assumptions within the algorithms. Also, a single clustering algorithm, for example, k-means, can produce different clusters when it is provided with data that differ in record length, number of dimensions (number of predictors), or initial conditions (e.g., starting centroid location). After discussing some of the similarities and differences in results produced by the three *base* clustering algorithms, methods to combine these clustering results are presented.

4.4.1 Clustering using k-means

The k-means algorithm is a commonly used clustering technique, where the goal is to identify subgroups within the data such that the data points belonging to a subgroup are similar to each other when compared to data points belonging to other subgroups. The Euclidean distance is used to determine the similarity of data points. The k-means algorithm was described in section 4.3.1.

Figure 4.7 shows the homogeneous drought regions obtained when considering 12-month droughts ending in May using SPI. The period of analysis was 1901-2004. The three panels in Figure 4.7 correspond to the following

(a) SPI: Only 12-month SPI drought characteristics were used as the predictor set. Therefore, the initial dimensionality was 104, corresponding to the number of years of record. The dimensionality was reduced using principal component analysis (PCA), and only 47 principal components (PC) that explained at least 90 percent of the variance were used as input to the k-means algorithm. As k-means does not automatically provide the optimal number of clusters, the algorithm was tested with K values ranging from 2 to 12. For each K value, several similarity statistics such as the Davies-Bouldin index (DB), Silhouette coefficient, and BIC were calculated. The optimal number of clusters for the given input data can be decided using any of these similarity metrics. In most cases, these indices report the same number for optimal clusters; however, sometimes, they were found to differ slightly. For this study, optimal number of clusters were reported according to the Davies-Bouldin index. Accordingly, k-means with $K = 9$ was found to be optimal with a DB value of 1.91.

(b) SPI-LL: In addition to 104 years of 12-month SPI drought characteristics ending in May, latitude and longitude information for each grid point were also included as predictors. The values of predictors were normalized before clustering. A total of 10 Homogeneous clusters were obtained with a DB value of 1.87.

(c) SPI-LL-LU: Finally, the percentage of land use under 19 land use classes listed in Table 4.1 were calculated over $1^\circ \times 1^\circ$ grids, the values were normalized and used as additional predictors to the k-means algorithm. Ten clusters with a DB value of 1.92 were found to be optimal.

The number of homogeneous clusters and their shape was found to vary slightly depending on the number of predictors used as input to the k-means algorithm. Therefore, even though the same drought index was used (SPI), the cluster assignment was found to vary significantly when additional information was incorporated in the form of new predictors (e.g., geographic information and land use). However, the cluster shapes formed were found to be mostly similar (92 percent of the grids), when comparing only Fig. 4.7b and Fig. 4.7c, indicating that land use information provided very little additional information to the k-means algorithm.

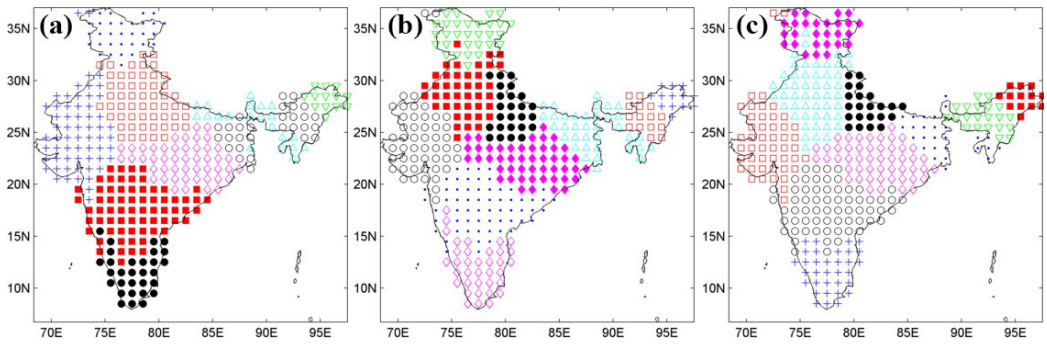

Figure 4.7. Homogeneous drought regions corresponding to 12-month time window ending in May for the study period 1901-2004 (P4) using k-means and the following predictors (a) SPI, (b) SPI-LL, and (c) SPI-LL-LU.

Figure 4.8 shows the homogeneous drought regions over India, using k-means algorithm on SPI time series computed at high spatial resolution ($0.25^\circ \times 0.25^\circ$). The time window of analysis was 12-months, and the entire period of record of high-resolution data (1901-2019) was used in the analysis. When using only SPI values as the input, nine clusters with a DB value of 2.4 were found to be optimal (see Figure 4.8a). When SPI values and geographic data were used together as inputs,

nine clusters with a DB value of 2.38 were optimal (Figure 4.8b; SPI-LL). Similarly, Figure 4.8c shows eight homogenous drought regions when SPI values in combination with geographic information and land use characteristics (i.e., SPI-LL-LU) were used as inputs to the k-means algorithm. Results indicate that the shape of the clusters was significantly different compared to those obtained using lower resolution ($1° \times 1°$) data (see Figure 4.7). Further, when using high resolution data, geographic information added little information as the regions in Figures 4.8a and 4.8b are similar. Also, the study found that the number of clusters and their shapes was similar when the period of analysis was 104 years (i.e., 1901-2004) instead of 119 years (i.e., 1901-2019).The choice of data resolution used in regionalization can thus lead to different conclusions. The scale-dependence indicates that the spatial resolution of input datasets have to be improved until the results do not change. Higher-resolution datasets may be useful for certain applications where getting precise boundaries is important, and studies over smaller regions.

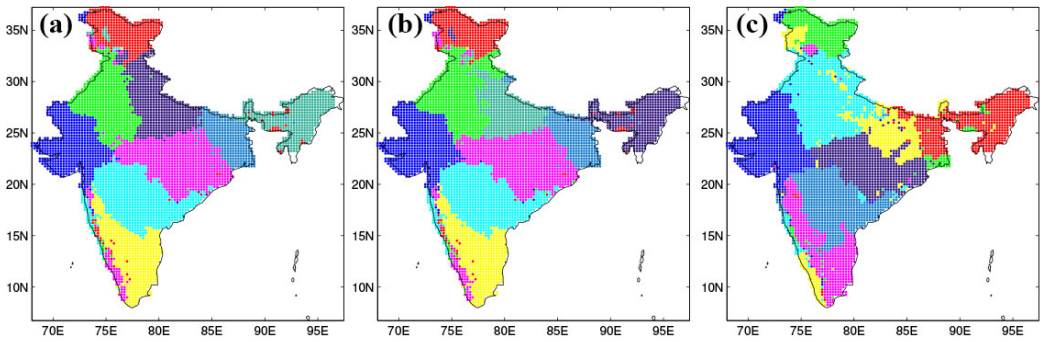

Figure 4.8. Homogeneous drought regions corresponding to a 12-month time window ending in May using k-means and the following predictors (a) SPI, (b) SPI-LL, and (c) SPI-LL-LU. The SPI values were computed using $0.25° \times 0.25°$ IMD precipitation dataset. The study period was 1901-2019.

Figure 4.9 shows the homogeneous drought clusters over India, when using the same clustering algorithm (k-means), period of analysis (1901 to 2004), geographic information and land use characteristics, but different drought indices. Specifically,

ten clusters with a DB value of 1.92 were found to be optimal when using SPI as the drought index (Figure 4.9a; SPI-LL-LU). Similarly, nine clusters were found to be optimal with a DB value of 1.74 for SPEI (Figure 4.9b; SPEI-LL-LU), seven clusters with a DB value of 2.44 were optimal for probabilistic SPI (Figure 4.9c; pSPI-LL-LU), and eleven clusters with DB value of 1.84 were optimal when drought characteristics pooled from SPI, SPEI, and pSPI, along with geographic and land use information were used as inputs to k-means algorithm (Figure 4.9d; Combined-LL-LU).

The number of optimal clusters formed was found to change with the choice of drought index. However, it should be noted that the choice of drought index also changes the dimensionality of the predictor set. For example, when using probabilistic SPI, the number of predictors before dimensionality reduction is 333, as it includes the probability of droughts to be in extreme, severe, and moderate drought category for 104 years, two geographic indicators, and 19 land use indicators. Similarly, when combining drought characteristics from the three drought indices, the dimensionality of the predictor set before PCA is 541. Therefore, the chosen drought index and the resulting increase (or decrease) in the dimensionality of predictors can lead to a significantly different number and shape of homogeneous drought clusters.

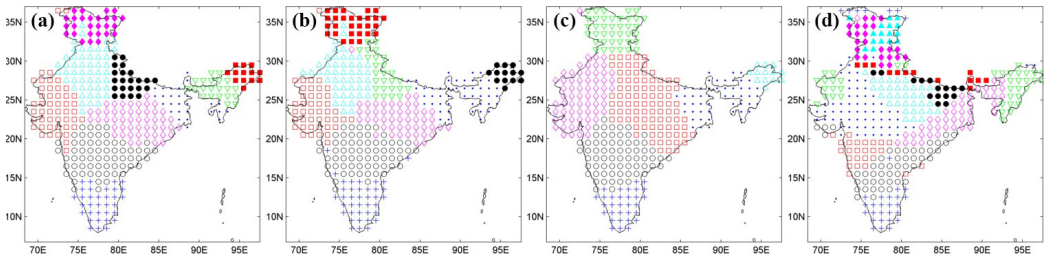

Figure 4.9. Homogeneous drought regions corresponding to 12-month time window ending in May for the study period 1901-2004 (P4) using k-means and the following predictors (a) SPI-LL-LU, (b) SPEI-LL-LU, (c) pSPI-LL-LU, and (d) Combined-LL-LU.

The sensitivity of drought clusters to the size of the study domain was analyzed. IMD grids over Jammu, Kashmir, and northeastern states were excluded from the

analysis, and only those belonging to the core-monsoon region of India were used as inputs. Figure 4.10 shows the homogenous clusters identified over the core-monsoon region using k-means algorithm and SPI-LL-LU, SPEI-LL-LU, pSPI-LL-LU, and SPI-LL-LU datasets as the inputs. In the case of SPI-LL-LU, ten clusters with a DB value of 2.09 was found to be optimal (Figure 4.10a). Similarly, nine clusters with a DB value of 1.91 for SPEI (SPEI-LL-LU; Figure 4.10b), eight clusters with a DB value of 2.87 for probabilistic SPI (pSPI-LL-LU; Figure 4.10c), and ten clusters with a DB value of 1.94 were found to be optimal when drought characteristics from all three indices were combined as inputs to the *k-means* clustering algorithm (Combined-LL-LU; Figure 4.10d). The number of *optimal* clusters, and their shape were found to be different compared to clusters obtained using data over the entire study domain (see Figure 4.9). Figure 4.11 shows the sub-optimal solutions for SPI-LL-LU ($k = 7$), SPEI-LL-LU ($k = 7$), pSPI-LL-LU ($k = 6$). The shape of these clusters were similar, except for minor differences, to the optimal clusters obtained for the full domain (Figure 4.9). The similarity between clusters in Figures 4.9a-c and Figures 4.11a-c were quantified using the adjusted Rand index (Hubert and Arabie, 1985). The adjusted Rand index for SPI-LL-LU, SPEI-LL-LU, and pSPI-LL-LU were 0.95, 0.86, and 0.64, respectively, indicating high similarity as these values are closer to one. However, similar conclusions could not be drawn when using Combined-LL-LU inputs. The highest adjusted Rand index was 0.3 (i.e., closer to zero, indicating dissimilar clusters) when the number of sub-optimal clusters was six. The high-dimensionality of input datasets and inclusion of relatively less reliable precipitation and land use data over Jammu, Kashmir, and northeastern states may have resulted in spatially less-compact clusters (Figure 4.9d), compared to Figure 4.10d, leading to the overall dissimilarity in results for Combined-LL-LU.

Next, the sensitivity of homogeneous clusters to the choice of period of analysis was investigated. A total of four different record lengths were used as inputs. The full data record of 104 years, 1901-2004 (P4), was further divided into three sub-periods 1901-1935 (P1), 1936-1970 (P2), and 1971-2004 (P3) similar to Chapter 2 (Mallya

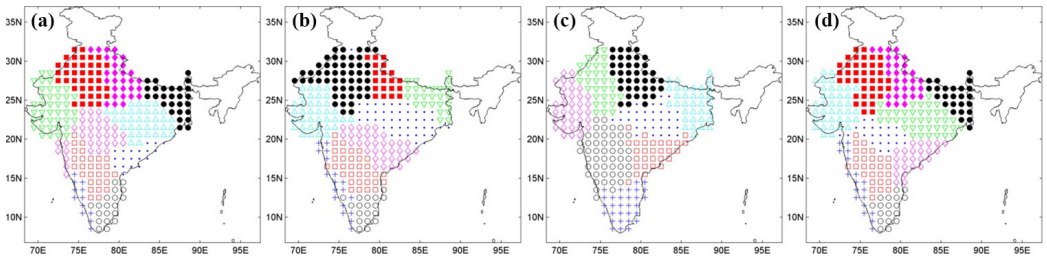

Figure 4.10. Homogeneous drought clusters over core-monsoon region of India. Results correspond to 12-month time window droughts ending in May during the study period 1901-2004 (P4) using k-means and the following predictors (a) SPI-LL-LU, (b) SPEI-LL-LU, (c) pSPI-LL-LU, and (d) Combined-LL-LU.

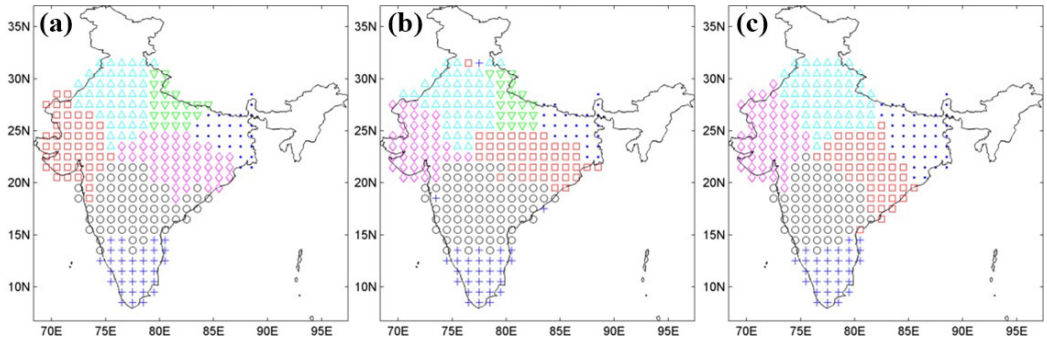

Figure 4.11. Sub-optimal homogeneous drought clusters over core-monsoon region of India. Results correspond to 12-month time window droughts ending in May during the study period 1901-2004 (P4) using k-means and the following predictors (a) SPI-LL-LU, (b) SPEI-LL-LU, and (c) pSPI-LL-LU.

et al., 2016). For brevity, results presented below are for drought characteristics obtained from SPI and includes geographic and land use information (SPI-LL-LU). For the period 1901-1935 (P1), the number of homogeneous clusters according to the k-means algorithm was 9 with a DB value of 1.83 (Figure 4.12a). For the period 1936-1970 (P2), nine clusters with DB value of 1.98 were found to be optimal (Figure 4.12b). Though the number of clusters for periods P1 and P2 were similar, the shape and cluster assignments were different. Note that the record length for these two

periods is the same (35 years), and the only difference is the drought characteristics. Therefore, the choice of period of analysis plays an important role in determining the size and location of homogeneous drought regions. For the period of 1970-2004 (P3) eight clusters with a DB value of 2.0 was found to be optimal (Figure 4.12c). Figure 4.12d corresponds to the full record period of 1901-2004 (P4), and has 10 clusters with a DB value of 1.92. The changes in the cluster size and shape were prominent in the Central Northeast region that includes the Indo-Gangetic plain - an agriculturally intensive region over India. Thus, from a policy-making and management perspective, it is important to continuously update the homogeneous drought clusters for effective planning and mitigation of droughts.

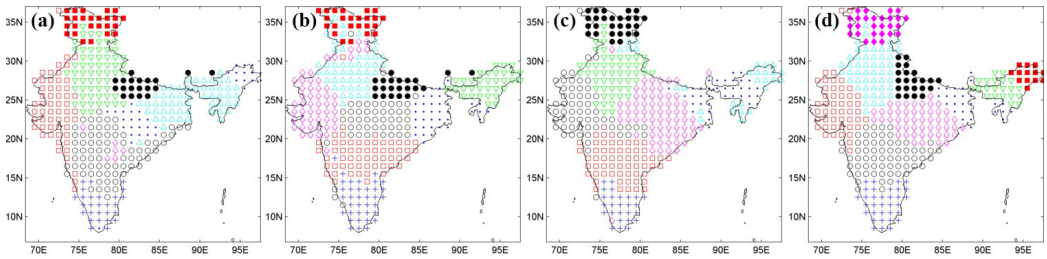

Figure 4.12. Homogeneous drought regions corresponding to 12-month time window ending in May using k-means and SPI-LL-LU predictor set for the periods (a) 1901-1935 (P1), (b) 1936-1970 (P2), (c) 1971-2004 (P3), and (d) 1901-2004 (P4).

Figure 4.13 shows the sensitivity of homogeneous drought clusters to the choice of threshold during dimensionality reduction in PCA. The dimensionality reduction is based on a user-defined threshold, *percentage of variance explained*, and was found to play a role in the eventual formation of clusters. Here, the initial predictor set was kept consistent (119 dimensions), i.e., SPI drought characteristics for a 12-month time window ending in May for the period 1901-2004, latitude, longitude, and land use characteristics. The only variable was the user-defined threshold during dimensionality reduction using PCA. Figure 4.13a shows that the number of clusters formed was 9, with a DB value of 1.55 when the variance-explained threshold was set to 70

percent. With this threshold, the dimensionality of the input dataset was reduced to 22. Figure 4.13b corresponds to a user-defined threshold of 80 percent for variance-explained while choosing principal components following PCA - resulting in reducing the number of dimensions to 34. This resulted in the formation of 10 clusters with a DB value of 1.74. Figure 4.13c shows the cluster formation when the variance-explained threshold was set to 90 percent during PCA, leading to 10 clusters with DB value of 1.92. While the choice of the threshold had an effect on the number of clusters for variance-explained values of 70 percent and 80 percent, there was minimal difference when comparing results for threshold choice of 80 percent and 90 percent.

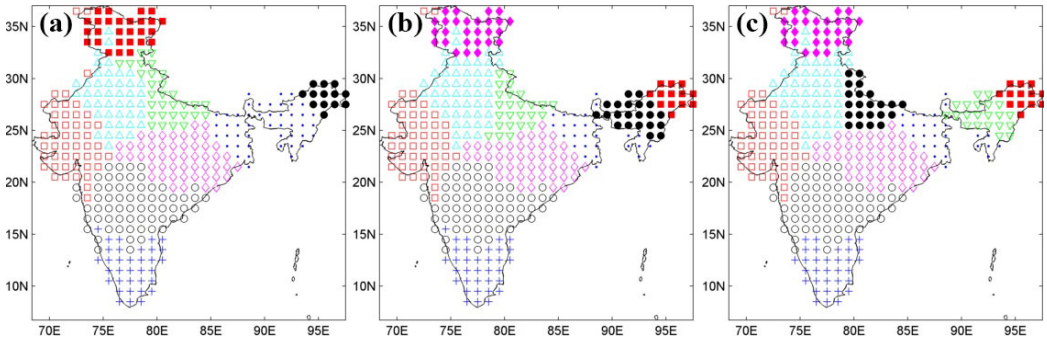

Figure 4.13. Homogeneous drought regions corresponding to 12-month time window ending in May for the study period 1901-2004 (P4) using k-means and SPI-LL-LU but with following number of principal components as predictors (a) 22 (70 percent variance-explained threshold), (b) 34 (80 percent), and (c) 54 (90 percent).

The sensitivity of drought cluster to the choice of time-window of analysis was investigated. For this purpose, SPI drought characteristics for a 4-month time window ending in September (corresponding to the southwest monsoon months over India) were used along with latitude, longitude, and land use information as input variables. After dimensionality reduction using PCA, the k-means algorithm was used to obtain homogeneous drought clusters for different periods. Figures 4.14a and b compare the drought clusters for the period 1901-1935 (P1) and 1936-1970 (P2). These two periods have the same record length, but due to unique drought characteristics during each of

these periods, the resulting drought clusters are markedly different. The number of clusters formed during the periods P1 and P2 was 11 (DB value is 1.76) and 10 (DB value is 1.9), respectively. In addition to some clusters (grids in the southern-most part of India) being not geographically contiguous, the shape and size of clusters are different for both these periods. This indicates that for these two periods (P1 and P2), the 4-month drought characteristics ending in September were different, as well as the inability of the algorithm to find geographically contiguous clusters. The lack of geographic contiguity may be due to (a) lack of significant difference in drought characteristics for several grids at 4-month time scale and (b) shorter record length ($\sim$35 years).

A total of 8 drought clusters (DB value of 2.05) was found to be optimum for 4-month droughts ending in September during the period 1971 to 2004 (P3, see Figure 4.14c). Due to the smaller number of clusters, the homogeneous regions are relatively contiguous for this period. Figure 4.14d shows the drought clusters formed for 4-month time window droughts ending in September when considering full record length of 104 years (1901-2004, P4). The total number of clusters was 12, with a DB value of 1.96. The formed clusters were compact, except for a few grids. The cluster shapes and sizes were markedly different when compared to shorter period lengths. The number of clusters and their shapes were also found to be different when compared to those obtained from 12-month time window droughts (Figure 4.12). This indicates that the choice of the time window of droughts can have a significant influence on the formed regions, and therefore has to be accounted for by drought managers when planning adaptation measures for short-term droughts.

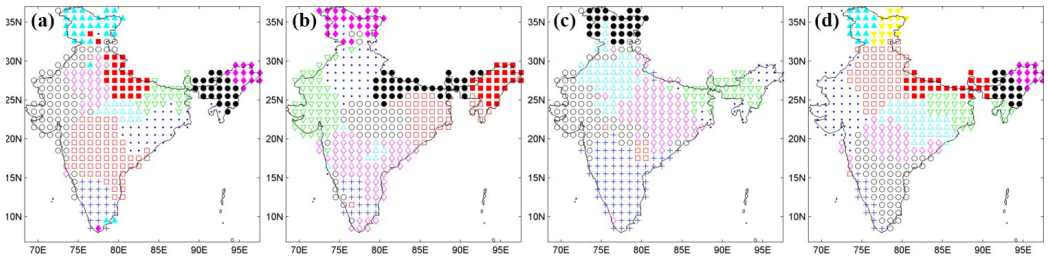

Figure 4.14. Homogeneous drought regions corresponding to 4-month time window ending in September using k-means and SPI-LL-LU predictor set for the periods (a) 1901-1935 (P1), (b) 1936-1970 (P2), (c) 1971-2004 (P3), and (d) 1901-2004 (P4).

4.4.2 Clustering using graph cuts

Next, the graph cuts algorithm (Boykov et al., 2001) described in Section 4.3.2 was used to find homogeneous drought regions over India (Figure 4.15). Drought characteristics for a 12-month time window ending in May using SPI, SPEI, pSPI, and combination of these three indices, geographic information, and land use classes were used as predictor variables. The period of analysis was set to 1901-2004 (P4). Figure 4.15a-c shows that the clusters formed in the peninsular region is similar for SPI, SPEI, and pSPI. Further, graph cuts method produced seven drought clusters with a DB value of 2.24 for SPI, nine clusters with DB value of 2.33 for SPEI, and nine clusters with DB value of 2.33 for pSPI. Eleven clusters with DB value of 1.89 were obtained when drought characteristics from all three drought indices were used as inputs to the model (Figure 4.15d; Combined-LL-LU). However, the clusters formed using the Combined-LL-LU predictor set were not compact over some regions - for example, south India, Orissa, and hilly regions over India (i.e., Jammu & Kashmir and NE states). The number and shape of the clusters obtained from graph cuts were different when compared to k-means clustering (Figure 4.9a-c-d) for SPI-LL-LU, pSPI-LL-LU, and Combined-LL-LU. However, the clusters are almost identical

in the case of SPEI-LL-LU. Thus, the choice of clustering algorithm also plays an important role in identifying homogeneous drought regions.

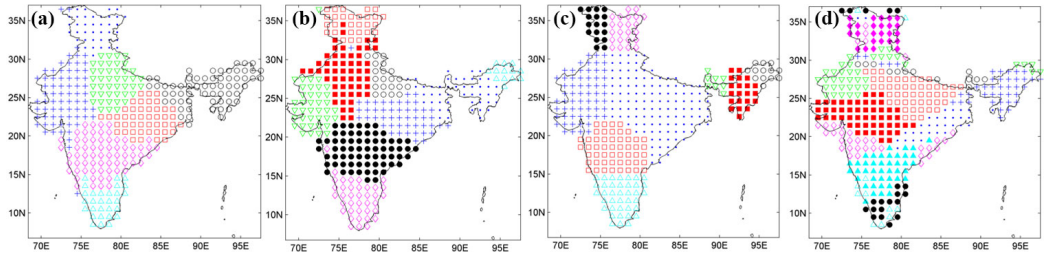

Figure 4.15. Homogeneous drought regions corresponding to 12-month time window ending in May for the study period 1901-2004 (P4) using graph cuts and the following predictors (a) SPI-LL-LU, (b) SPEI-LL-LU, (c) pSPI-LL-LU, and (d) Combined-LL-LU.

4.4.3 Agglomerative clustering

Agglomerative clustering is a hierarchical grouping technique where each grid point initially belongs to its own cluster. Pairs of grids are then merged together based on user-defined linkage criteria and a distance metric that decides how the final clusters are formed. The complete (or maximum) linkage criteria with Euclidean distance were chosen for this study. Figure 4.16a-d shows the homogeneous drought regions obtained from agglomerative clustering using four predictor sets SPI-LL-LU, SPEI-LL-LU, pSPI-LL-LU, and Combined-LL-LU, respectively. The time-window of 12-months ending in May and the period between 1901-2004 were chosen for this analysis. The number and shape of homogeneous clusters formed using this method were found to be different when compared to k-means clustering (Figure 4.9) and graph cuts clustering (Figure 4.15) methods. Using agglomerative clustering, eight clusters with a DB value of 2.17 were formed for SPI-LL-LU (Figure 4.16a). Similarly, nine clusters with a DB value of 1.98 were obtained for SPEI-LL-LU. Eight clusters with a DB value of 2.67 for pSPI-LL-LU and nine clusters with a DB value of 2.19

were formed for Combined-LL-LU. By using a tree-based or hierarchical clustering technique, the optimal cluster number and shapes were found to be different compared to algorithms that are based on cuts or partition of vertices in a graph (graph cuts) or to algorithms that assume the variance of the distribution of each attribute is spherical (k-means).

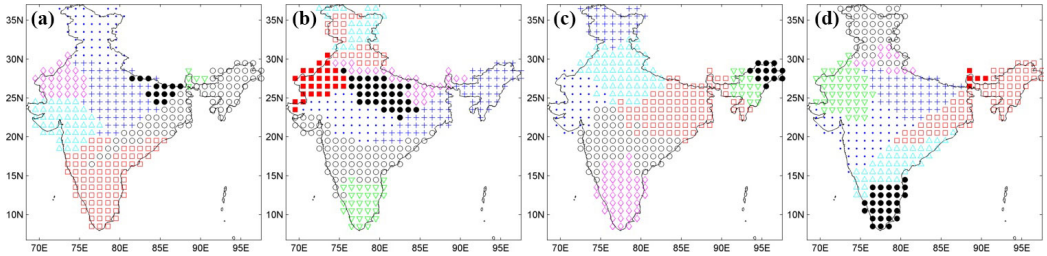

Figure 4.16. Homogeneous drought regions corresponding to 12-month time window ending in May for the study period 1901-2004 (P4) using agglomerative clustering and the following predictors (a) SPI-LL-LU, (b) SPEI-LL-LU, (b) pSPI-LL-LU, and (c) Combined-LL-LU.

4.4.4 Clustering of clusters

The number and shape of homogeneous drought clusters obtained using three different base clustering algorithms were presented in the preceding sections. The resulting clusters were found to be sensitive to multiple factors such as choice of the clustering algorithm, drought index, time window of drought, period of analysis, number of predictors in the input dataset, and model parameters of clustering algorithms. This can pose a challenge to water resources managers and policy-makers as no objective methodology exists or is used in practice to combine results of regionalization, especially in the context of droughts.

Figure 4.17 shows the results of combining clustering results obtained from base algorithms, namely, k-means, graph cuts, and agglomerative clustering. A pairwise similarity matrix is first constructed between pairs of the base clustering algorithms to

achieve the objective of combining base clustering results. These similarity matrices are then combined to get a consensus matrix using equation 4.14 (Monti et al., 2003). Finally, the consensus matrix is used as an input to a hierarchical algorithm with complete linkage criteria to obtain the final clustering. The final clusters are based on the similarity of cluster shapes and structure obtained from base clustering methods. Figure 4.17a shows the clustering of clusters for SPI-LL-LU dataset, and has nine homogeneous clusters with a DB value of 2.34. The base clusters were obtained considering drought characteristics for a 12-month time window corresponding to the period 1901 to 2004. Similarly, Figure 4.17b-c-d shows nine clusters each for SPEI-LL-LU, pSPI-LL-LU, and Combined-LL-LU predictor sets with DB values of 1.8, 2.22, and 1.99, respectively. The shape and size of the clusters vary significantly for the four drought indices and therefore highlights the importance of choice of drought index for regionalization. However, it is possible to extend the methodology to obtain one unique final clustering result by combining results from twelve individual base clustering results discussed in the earlier sections (Figures 4.9, 4.15, and 4.16).

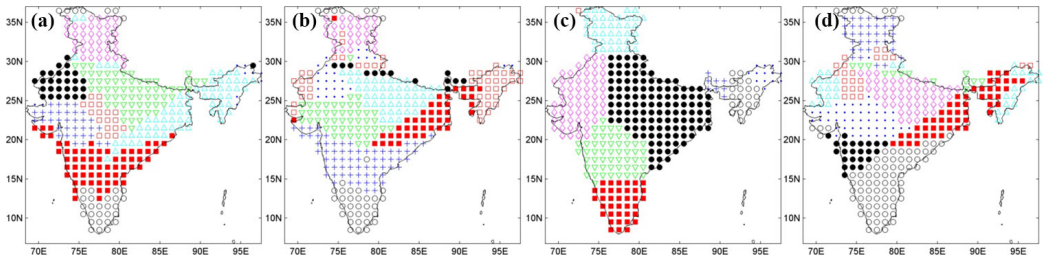

Figure 4.17. Clustering of homogeneous drought clusters using similarity matrix and complete linkage criteria for (a) SPI-LL-LU, (b) SPEI-LL-LU, (c) pSPI-LL-LU, and (d) Combined-LL-LU predictor sets. The base clusters from k-means, graph cuts, and agglomerative clustering with optimal Davies-Bouldin scores were chosen.

In figure 4.18, the combined clusters obtained using connected triple-based similarity matrix and complete linkage criteria (CTS-CL) is presented. The base clusters were obtained using k-means, graph cuts, and agglomerative algorithms and corre-

spond to a 12-month drought time window over the period 1901 to 2004. In this approach, weighted graphs are constructed by comparing cluster assignments of input data according to multiple base clustering algorithms. These weighted similarity matrices are then used to obtain the final clustering using a hierarchical method with complete linkage criteria. Figures 4.18a-b, corresponding to SPI-LL-LU and SPEI-LL-LU datasets, show that the combined clusters have more similarity in their geographical extent than differences. The combined clusters for both datasets have nine homogeneous regions with DB values of 1.85 and 1.64.

The clustering of base clusters produced six homogeneous regions when considering pSPI-LL-LU dataset with DB value of 2.58 (Figure 4.18c). Similarly, nine clusters were obtained when considering Combined-LL-LU with DB value of 1.86. The shape and size of the homogeneous regions vary significantly when comparing results for SPI-LL-LU and SPEI-LL-LU (Figure 4.18a-b) with pSPI-LL-LU and Combined-LL-LU (Figure 4.18c-d). This, once again, highlights the important role that the choice of drought index plays in regionalization.

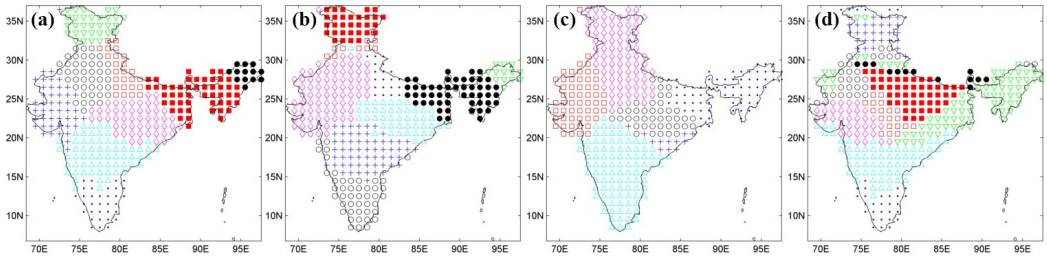

Figure 4.18. Clustering of homogeneous drought clusters using CTS-CL for (a) SPI-LL-LU, (b) SPEI-LL-LU, (c) pSPI-LL-LU, and (d) Combined-LL-LU predictor sets. The clusters formed correspond to a drought time window of 12-months ending in May and for the period 1901-2004. k-means, graph cuts, and agglomerative algorithms were used to obtain the base clusters.

4.4.5 Regional Drought Characteristics

As described in Section 4.3.7, for a given drought index (e.g., SPI-4, SPI-12, SPEI-4, etc.) the drought characteristics such as the number of drought events, duration of each drought event, and average intensity of droughts during each drought event were calculated using run theory at each grid over the study area. The results presented below correspond to the SPI-12 drought index over the period 1901-2004. The drought characteristics were aggregated over homogeneous drought clusters (Figure 4.18a), and their box plots were analyzed. For different durations, average drought intensity values were aggregated to obtain an annual minimum series. Probability distributions that provide the best fit according to Chi-square goodness of fit test were identified. The GEV distribution was found to provide the best fit. Frequency analysis was then performed to obtain average drought intensity values for different return periods. The areal extent of droughts and the average drought intensity over drought-affected grids were analyzed over different homogeneous drought regions during the period 1901-2004. Frequency analysis was performed to obtain drought intensity values for different return periods and areal extent of droughts. Finally, over each homogeneous drought region, the available precipitation data were pooled from all grids over a region to obtain a representative time series. The representative series was later used to analyze historic droughts and their spatial extent from a regional drought-management perspective.

4.4.5.1 Characteristics of drought events

Drought characteristics such as average drought intensity during a drought event, the duration of drought event in months, and the number of drought events were computed for SPI-12 series during the period 1901-2004 using moving 12-month windows (i.e., SPI-12 values for 1248 months) at all grids over the study area. These drought characteristics are compared over homogeneous drought regions (refer to Figure 4.21 for geographic context) using box plots (Figure 4.19). Figure 4.19a indicates that the

median value of average intensity of SPI-12 drought events is between -1.0 and -1.5 across regions. The width of the box plot varies across different regions, indicating a wide range of possible average drought intensity values. Figure 4.19b shows the box plots of the duration of SPI-12 drought events recorded at IMD grids over different regions. The median drought duration is between 3 to 9 months. The width of the box plot is more or less uniform across regions (8 out of 9). Several outliers indicate that some grids across regions experienced significantly long duration SPI-12 droughts (30 to 60 months) during the study period 1901-2004. Figure 4.19c shows the number of SPI-12 drought events recorded over IMD grids during the period 1901-2004 over different homogeneous regions. The median number of drought events for SPI-12 is between 40 to 55 (7 out of nine regions). The width of the box plots (along the y-axis) varies across all regions, indicating the differences in drought characteristics among the regions.

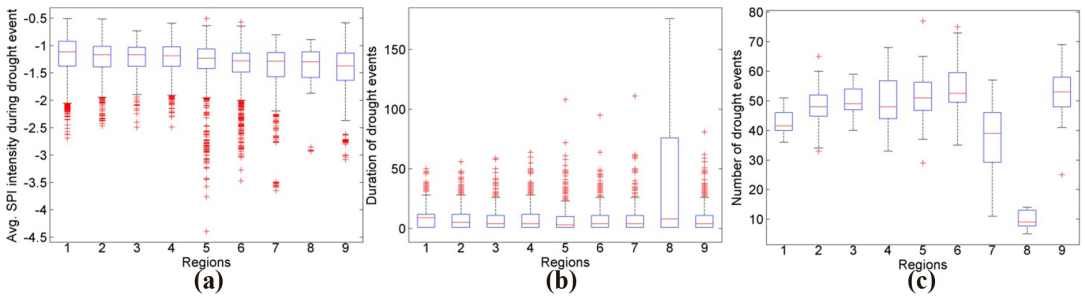

Figure 4.19. Variation of (a) average drought intensity during drought events, (b) duration of drought events, and (c) number of drought events across grids belonging to nine homogeneous regions over India considering SPI-12 drought series for the period 1901-2004. For geographic context of homogeneous regions see Figures 4.18a and 4.21.

For each grid over a homogeneous region and for moving 4-month window SPI series during the period 1901-2004, drought characteristics such as average drought intensity during a drought event, the duration of drought event in months, and the number of drought events over a grid were analyzed. The results for these drought characteristics are presented as box plots (Figure 4.20) for each homogeneous region

(refer to Figure 4.22 for geographic context). Figure 4.20a indicates that the median value of the average intensity of 4-month SPI drought events is between -1.0 and -1.5 across regions. The width of the box plot, indicating the distribution of average drought intensity values are similar across the regions. Also, the smaller width of the box plots indicates less variation of average drought intensity values during drought events across all grids over a given region. Figure 4.20b shows the box plots of the duration of SPI-4 drought events recorded at IMD grids over different regions. The median drought duration is two months. The width of the box plot is uniform for the majority of the regions (8 out of 9). However, it is important to note the presence of outliers, indicating that several grids across all regions experienced significantly long duration SPI-4 droughts (10 to 50 months) during the study period 1901-2004. Figure 4.20c shows the number of SPI-4 drought events recorded over IMD grids between 1901 and 2004 (total of 1248 months) over different homogeneous regions. The median number of drought events for 4-month SPI is above 130 (7 out of nine regions). The width of the box plots varies across all regions, indicating the differences in drought characteristics among the regions.

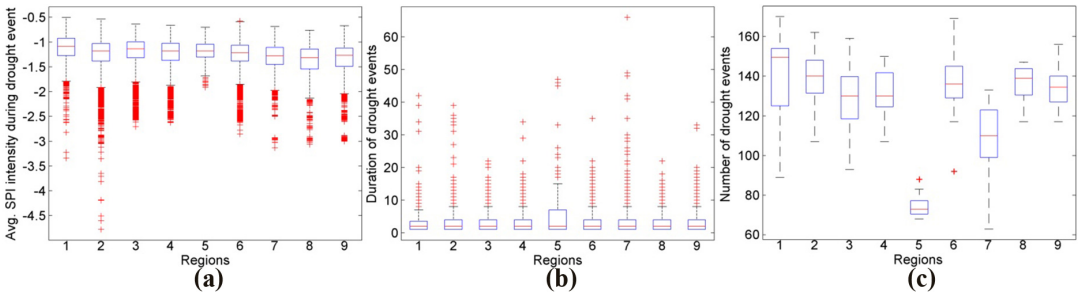

Figure 4.20. Variation of (a) average drought intensity during drought events, (b) duration of drought events, and (c) number of drought events across grids belonging to nine homogeneous regions over India considering SPI-4 drought series for the period 1901-2004. For geographic context of homogeneous regions see Figure 4.22.

4.4.5.2 Intensity-Duration-Frequency analysis

For each homogeneous region, SPI-12 drought events of the same duration were identified at all grids. For example, for homogeneous region-1 over Gujarat in western India (see inset in Figure 4.21a), the average SPI-12 drought intensities of droughts with 3-month duration were tabulated at all grids. Then, using previously tabulated values, for each year during the study period (1901-2004), the minimum value of the average SPI-12 drought intensities was recorded. Frequency analyses were carried out on this annual minimum series of average SPI-12 drought intensities. Among several distributions (Gumbel, GEV, t-location-scale, logistic, etc.), the GEV distribution provided the best fit according to the Chi-square goodness of fit test with a significance level (α) of 1%. Using the parameters of the best-fit distribution average SPI-12 intensities for drought events with a 3-month duration were estimated for 2, 5, 10, 25, 50, and 100 year return periods. This process was repeated for other durations of interest and for all homogeneous drought regions. In addition to providing insights into drought characteristics over an area, IDF plots are also useful in estimating return periods of any drought event. For example, if the average SPI-12 intensity of the 2002-2003 drought event of a 6-month duration is -2.5, then the return period can be estimated as 25 years (Figure 4.21a). In general, the average SPI-12 intensity of drought events with a duration of 6-months or higher is more severe compared to drought events of length less than 6-months. Also, for peninsular and northeastern parts of India (Figure 4.21e-f-i), that are among the highest rainfall receiving regions, rare drought events (T = 50 or 100 years) of longer duration ($>$ 6 months) tend to be severe compared to other regions.

Figure 4.22 shows the IDF curves developed for nine homogeneous regions of SPI-4. The geographic extents of the homogeneous clusters are shown in the inset of each subplot. As described above, over each region, frequency analyses were carried out on annual minimum series of average SPI-4 drought intensities of a selected duration event. The GEV distribution provided the best fit for the annual series according

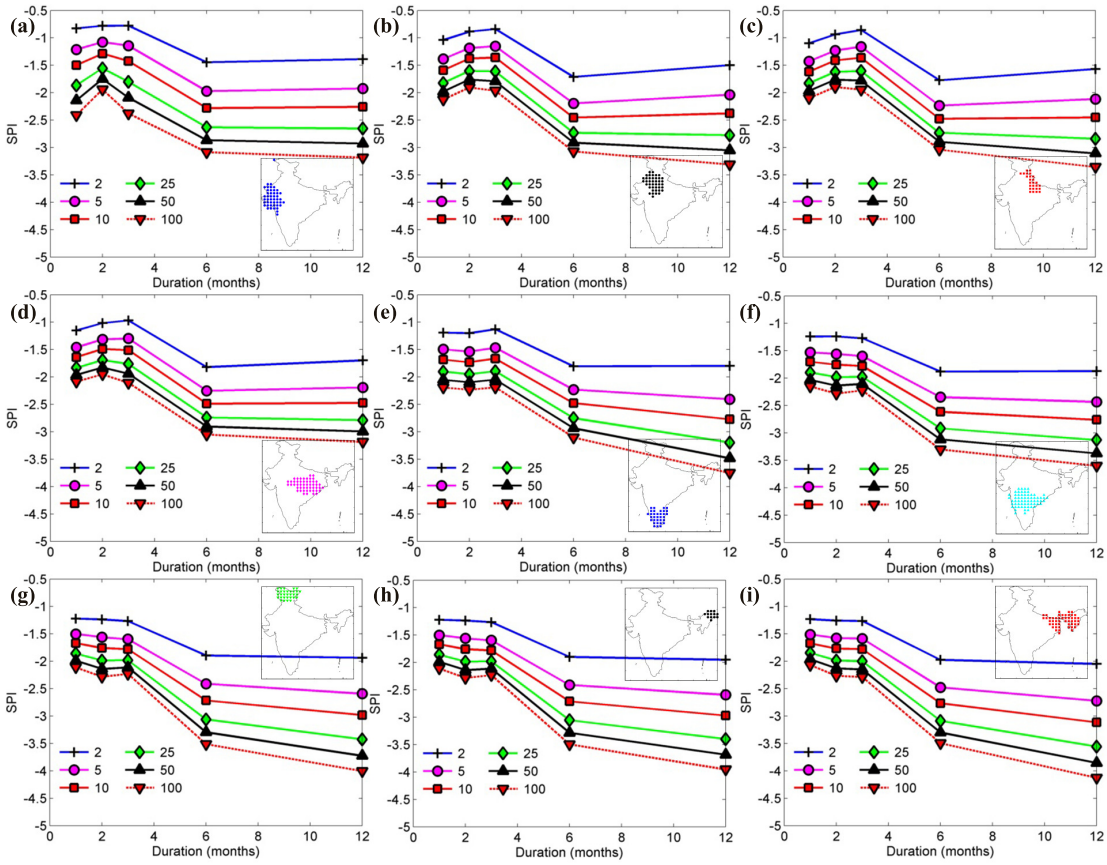

Figure 4.21. Intensity-Duration-Frequency curves of SPI-12 drought series over nine homogeneous regions (a) - (i) over India. The geographic extent of each homogeneous region is shown as an inset within the subplots. Each curve represents a return period of interest.

to the Chi-square goodness of fit test with a significance level (α) of 1%. Using the parameters of the best-fit distribution, average SPI-4 intensities for the selected duration were estimated for 2, 5, 10, 25, 50, and 100 year return periods. From Figure 4.22, it may be noted that the average SPI-4 intensity during drought events with a duration of 6-months or higher is more severe compared to shorter duration events. Also, over peninsular and western parts of India (Figure 4.22a-b-c), that receive majority of their precipitation during southwest monsoon season, less-frequent and rare drought events (T = 10, 25, 50 or 100 years) of longer duration (> 6 months) tend to be severe compared to other regions.

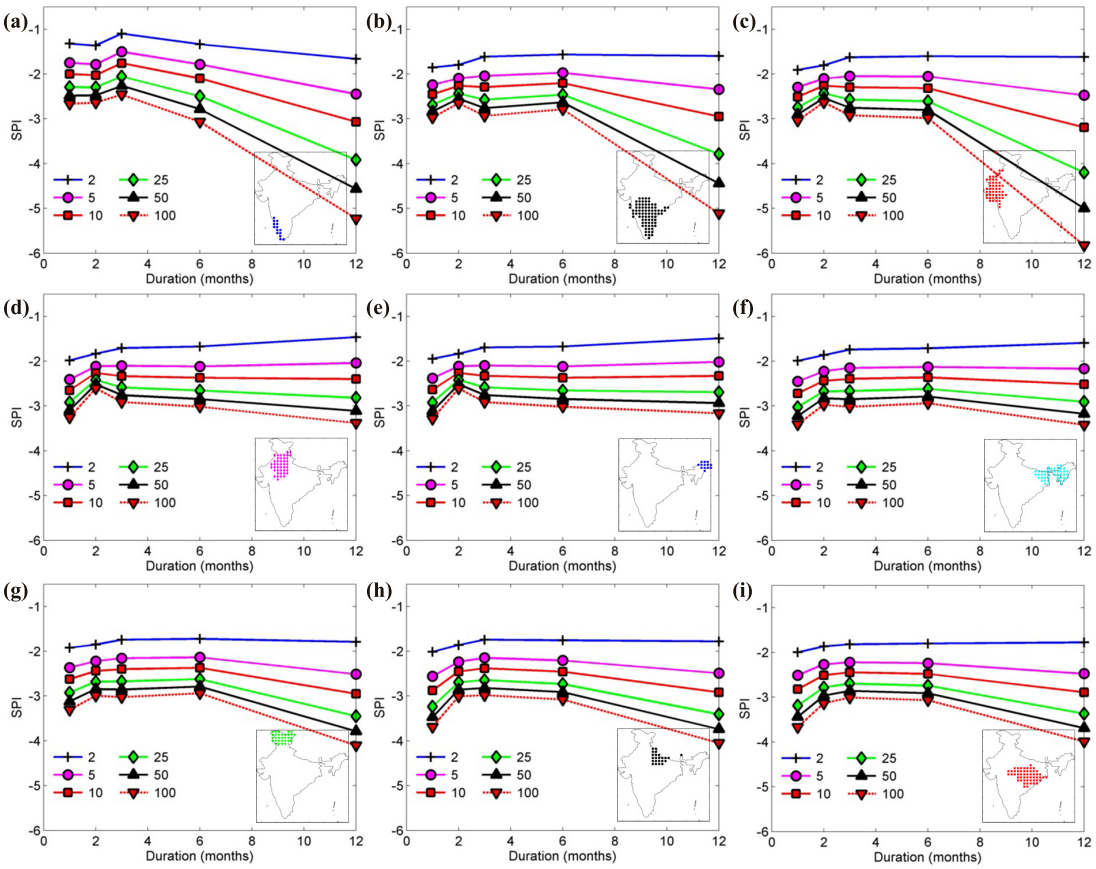

Figure 4.22. Intensity-Duration-Frequency curves of SPI-4 drought series over nine homogeneous regions (a) - (i) over India. The geographic extent of each homogeneous region is shown as an inset within the subplots. Each curve represents a return period of interest.

4.4.5.3 Areal extent of drought and average drought intensity

The homogeneous drought regions obtained after clustering of clusters (e.g., CTS-CL, Figure 4.18) were used to study the overall drought characteristics. The drought characteristics presented below correspond to SPI for a pre-selected time window (12-months, 4-months, etc.) for 1901-2004 and are for all months during the study period (i.e., using 12-months and 4-months moving window). For every month, the percentage of grids having SPI ≤ -1.0 was calculated to get the areal extent of drought over each region.

Figure 4.23a shows the box plot of areal extent of SPI-12 droughts over nine homogeneous regions over India (Figure 4.21). In each box plot, the central red line denotes the median value of the areal extent for a given region. The bottom and top edges of the box plot denote the 25^{th} and 75^{th} percentiles of the areal extent of droughts for a region, respectively. The whiskers of the box plot denote the range of values of the areal extent of droughts that were not considered outliers, and those that were outliers are shown using '+' symbol. According to Figure 4.23a, for the chosen threshold of SPI-12 ≤ -1.0, the median value of areal extent droughts lies between 10 to 15% for all regions except regions 7 and 8. The median areal extent of droughts in regions 7 and 8 (Jammu & Kashmir and northeastern states) is 26% and 100%, respectively. However, the width of the box plots indicates that during the study period, there were occasions when none or 50% and more grids belonging to all regions except region 8 were in a drought. Results over regions 7 and 8 that comprise of grids over Jammu & Kashmir and northeastern part of India (Figure 4.21g-h) are less reliable because the gridded IMD precipitation data over these regions were computed using sparse high-altitude rain gauge stations.

Figure 4.23b shows the distribution of the average intensity of SPI-12 droughts over drought-affected grids in each region. For each region, the average drought intensity was computed for each month of the study period considering only grids under drought (SPI-12 ≤ -1.0). The median SPI-12 drought intensity values were

around -1.35 for all regions. The spread of the whiskers of each box plot, ranging between -1.0 and -2.5, denotes the distribution of average SPI-12 intensity values over each region. The variation of the width of these box plots, and the number and magnitude of outliers indicates the difference in drought characteristics between the regions. Box plot for region 7, belonging to Jammu and Kashmir (Figure 4.21g), is the widest indicating a large range of average SPI-12 intensities compared to other regions. Also, box plots for regions 3, 5, 6, and 7 (see Figure 4.22c,e,f,i) show that they experienced extreme droughts (SPI-12 ≤ -3.0) on several occasions during the period 1901-2004.

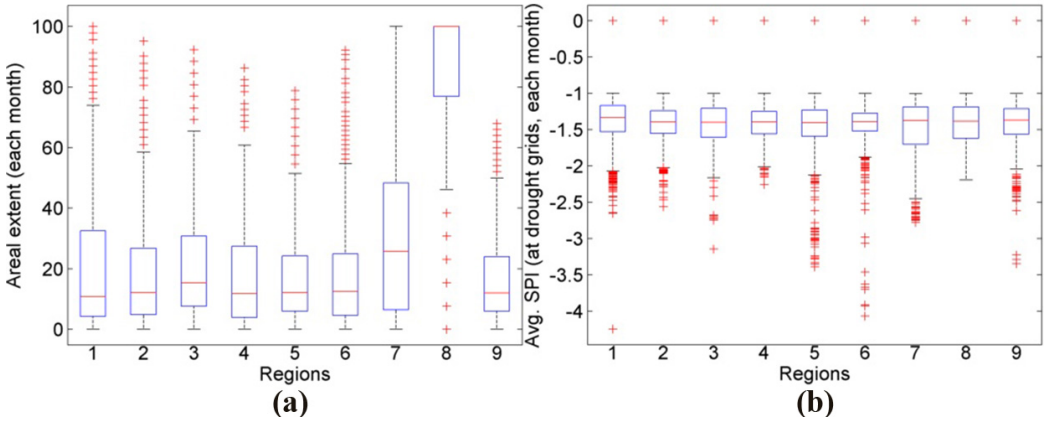

Figure 4.23. Comparison of (a) areal extent of droughts (%) and (b) average intensity of droughts across nine homogeneous drought regions using SPI-12 over 1248 months of the study period (1901-2004). For geographic context of homogeneous drought regions with respect to SPI-12-LL-LU see Figures 4.18a and 4.21.

Figure 4.24a shows the box plot of areal extent of SPI-4 droughts over nine homogeneous regions over India (see Figure 4.22 for geographic extent of each region). According to Figure 4.24a, the areal extent of droughts varies across regions. The lower limit is 0%, and the upper limit is 47% or more. The range of areal extent values indicates that during the study period, there were occasions when none to over 47% of grids belonging to different homogeneous regions experienced droughts based on

the chosen threshold (SPI-4 ≤ -1.0). The median areal extent was between 8 to 30 percent for all regions except region 5, which has a median value of 100%. Note that region five comprises of thirteen grids and is located in the northeastern part of India (Figure 4.22e). The median value of 100% for areal extent of SPI-4 droughts indicates that during more than half of the 1248 months of the study period (1901-2004), all grids over region 5 experienced SPI-4 ≤ -1.0.

Figure 4.24b shows the distribution of the average intensity of SPI-4 droughts over drought-affected grids in each region. The average drought intensity over a region was computed as described previously (SPI-4 ≤ -1.0 was chosen as the threshold). The median SPI-4 drought intensity values were around -1.4 for all regions. The spread of the whiskers (-1.0 to -2.2) of each box plot, denotes the distribution of average SPI-4 intensity values over each region. The variation of the width of box plots, as well as the number and magnitude of outliers indicates the difference in drought characteristics between the regions. Box plot for regions 1, 5, and 7, belonging to the western-coast, northeastern states, and Jammu & Kashmir region of India (Figure 4.22a,e,g), have wide widths indicating a large range of average SPI-4 intensities compared to other regions. Also, regions 2 and 6 (see Figure 4.22b,f) experienced extreme droughts (SPI-4 ≤ -4.0) on eight occasions during the period 1901-2004.

4.4.5.4 Intensity-Area-Frequency analysis

The spatial and temporal drought characteristics were used to develop Intensity-Area-Frequency (IAF) curves for each homogeneous drought region over India. For a pre-specified drought index and time window (e.g., SPI-12, SPI-4, etc.), the areal extent of drought over a region was tabulated for different drought intensity values using GIS (Loukas and Vasiliades, 2004). Frequency analysis was then performed by fitting a probability distribution to average drought intensities prevalent over a specific areal extent (say 10%). The parameters of the fitted distribution were then used to estimate average drought intensity values for several return periods of interest.

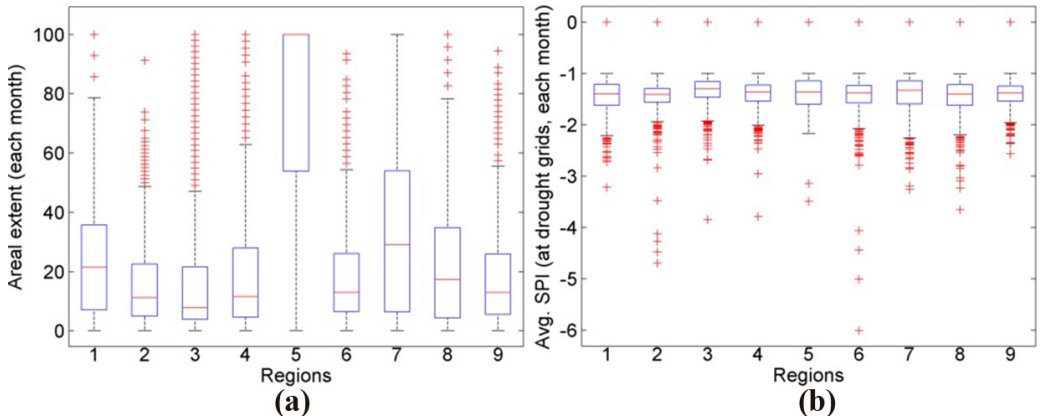

Figure 4.24. Comparison of (a) areal extent of droughts (%) and (b) average intensity of droughts across nine homogeneous drought regions using SPI-4 over 1248 months of the study period (1901-2004). For geographic context of homogeneous drought regions with respect to SPI-4-LL-LU see Figure 4.22.

Figure 4.25 shows the IAF curves at nine homogeneous drought regions over India, considering the SPI-12 drought index. The IAF curves are useful in determining the return period of drought events over an area if the areal extent and the average drought intensity values are known. For example, if the areal extent of drought over region 5 (Figure 4.25e, southern peninsular India) is 10%, and the average intensity of SPI-12 is -2.6, then it is a 10-year return period drought. Similarly, if the average SPI-12 is -2.6, and the areal extent is 40%, then the return period of drought is over 25 years, and so on. In general, one may conclude that for a fixed return period (say 10-years), the average SPI-12 intensity decreases with an increase in the areal extent. Tightly bound IAF curves indicate that even a slight increase or decrease in average SPI-12 intensity can lead to different conclusions about return periods (see Figure 4.25d).

Figure 4.26 shows the IAF curves at nine homogeneous drought regions over India, considering the SPI-4 drought index. As described before, the IAF curves are useful in determining the return period of drought events over a region. For example, if the

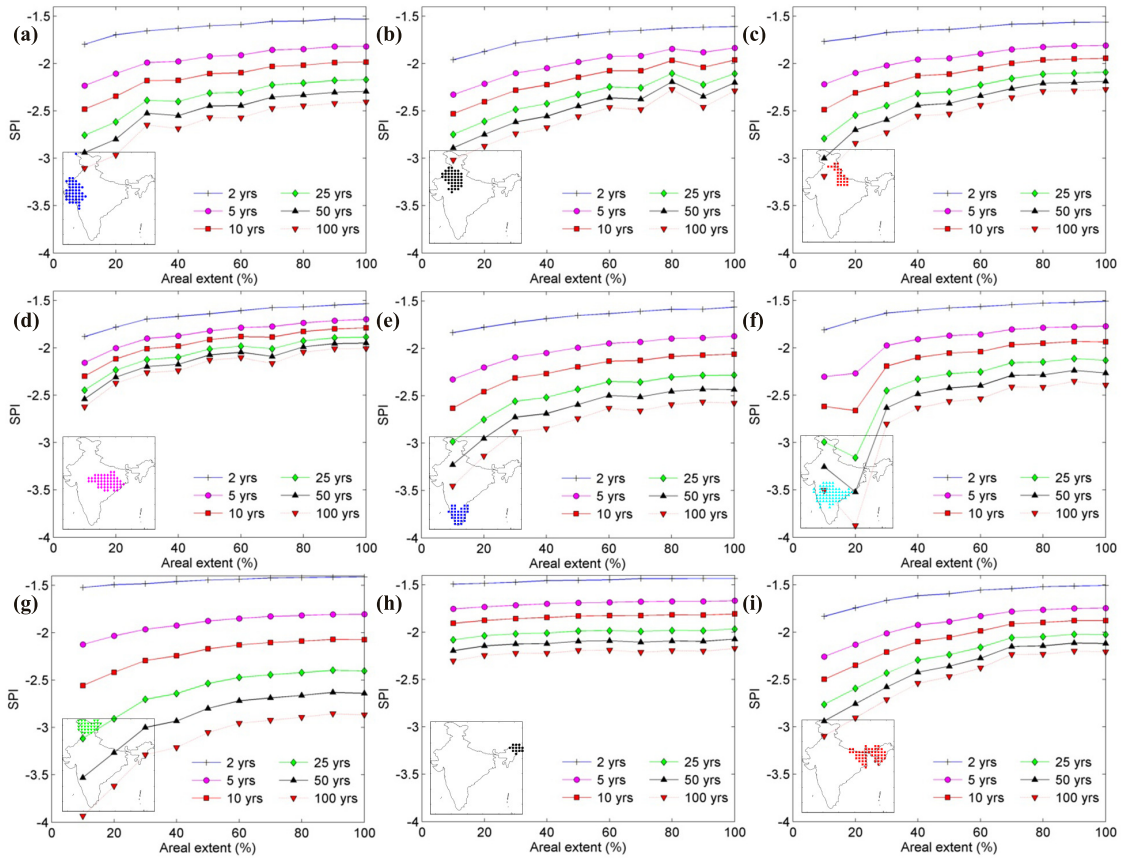

Figure 4.25. Intensity-Area-Frequency curves of SPI-12 drought series over nine homogeneous regions (a) - (i) over India. The geographic extent of each homogeneous region is shown as an inset within the subplots. Each curve represents a return period of interest.

areal extent of drought over region 1 (Figure 4.26a, western coast of India) is 10%, and the average intensity of SPI-4 is -2.4, then it is a 5-year return period drought. Now, if the areal extent increases to 30%, but the SPI-4 intensity remains unchanged, then such an event has a return period of 10 years, and so on. The general inverse relationship between drought intensity and areal extent for any fixed-return period holds true across all homogeneous drought regions for SPI-4. The IAF curves for SPI-4 were found to be tightly bound for larger return period events for most regions (e.g., Figure 4.26b,d,etc.).

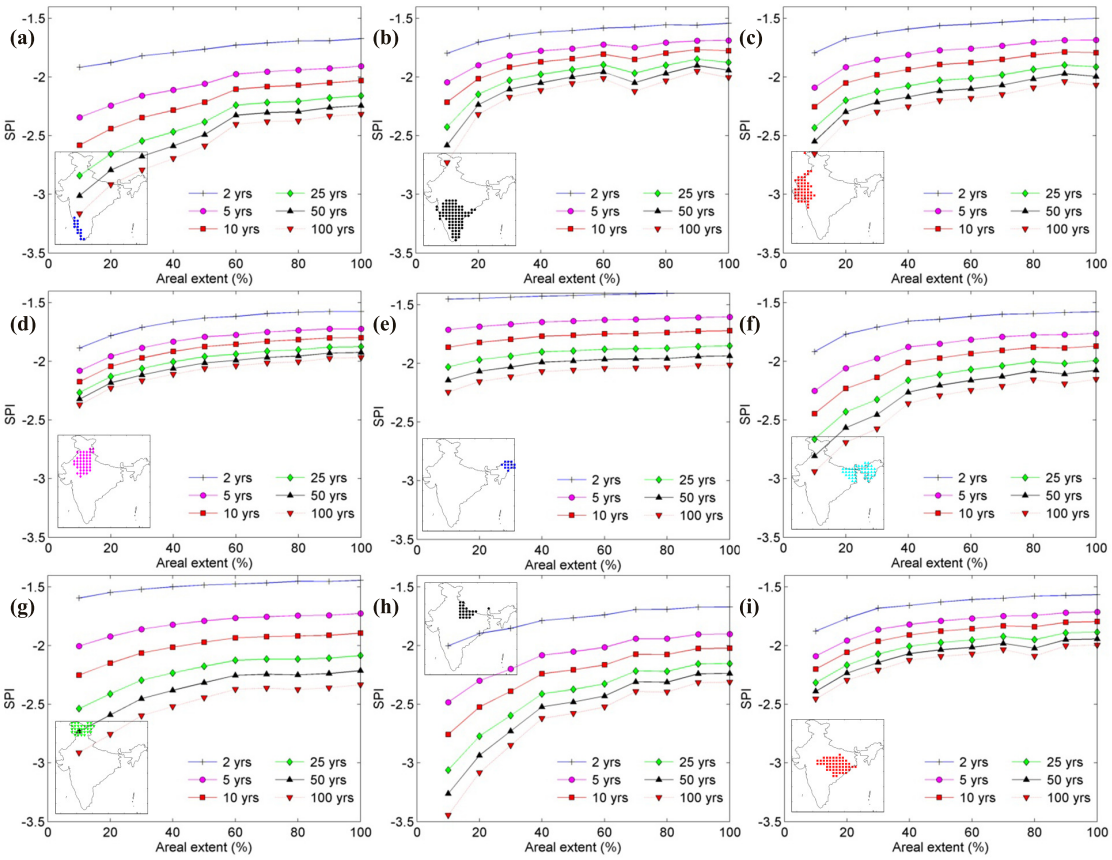

Figure 4.26. Intensity-Area-Frequency curves of SPI-4 drought series over nine homogeneous regions (a) - (i) over India. The geographic extent of each homogeneous region is shown as an inset within the subplots. Each curve represents a return period of interest.

4.4.5.5 Drought analysis by pooling data over homogeneous regions

The utility of identifying homogeneous drought regions over an area (e.g., Figure 4.18a) by providing a framework for regional drought analysis is presented below. First, a representative monthly precipitation time series was computed for each homogeneous drought region by pooling monthly precipitation data available at each grid within the region and then taking their averages. Then, the representative precipitation time series for each region was used to obtain SPI intensities for different time windows.

Figure 4.27 shows the SPI-12 intensities for ending month of May over the period 1902 to 2004 computed over nine SPI-12 homogeneous drought regions over India and shown as bar plots. The geographic extent of each region is shown as an inset within each sub-plot. Negative SPI values indicate drought conditions. For example, region 5 or southern peninsular India (Figure 4.27e) experienced extreme droughts during the years 1952, 2003, and 2004 with SPI-12 intensities of -2.09, -3.75, and -2.41, respectively. For the same years, while other regions also experienced drought conditions, their intensities differed. Also, for the year 2004, drought recovery was seen in most regions expect region 5. This regional drought perspective is useful for water resources planners in formulating region-specific water management plans.

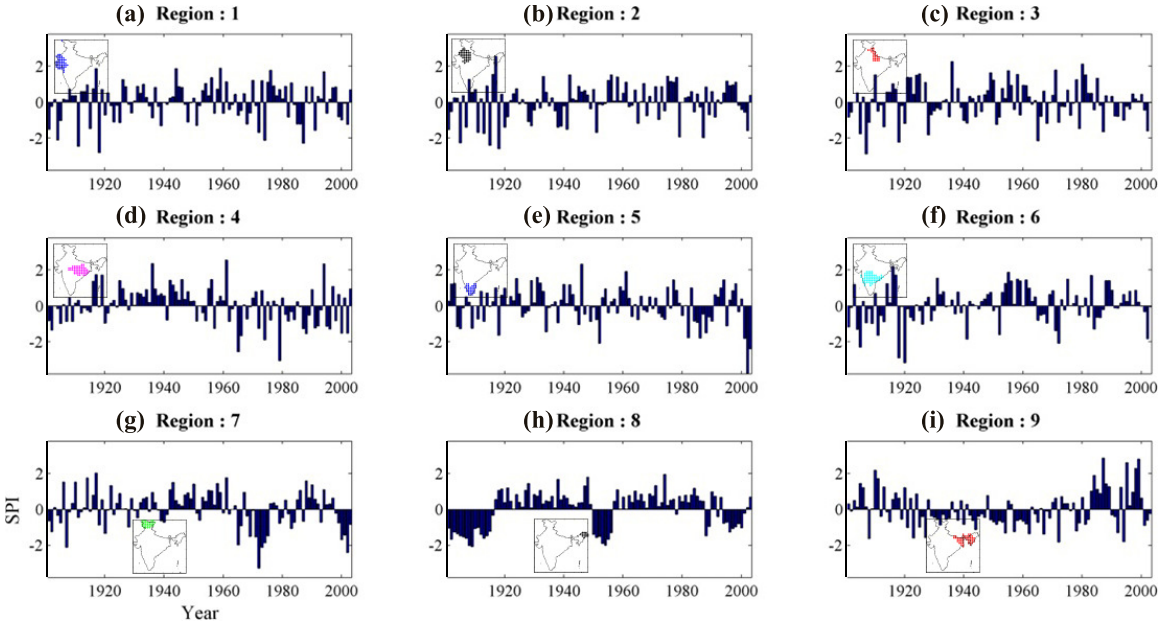

Figure 4.27. SPI-12 drought intensity ending in May for 1901-2004 computed after pooling precipitation data over nine homogeneous regions (a) - (i) over India. The geographic extent of each homogeneous region is shown as an inset within the subplots.

Figure 4.28 shows the areal extent of SPI-12 droughts ending in May at each of the nine SPI-12 homogeneous drought regions over India during the period 1902 to 2004. The geographic extent of each region is shown as an inset within each sub-

plot and the areal extent values (in percentage) are shown as bars. For any region, the areal extent values represent the percentage of grids in the region experiencing droughts (SPI ≤ -1.0) during May. For example, the areal extent of SPI-12 droughts in region 5 or southern peninsular India (Figure 4.28e) during the years 1952, 2003, and 2004 was 57.6%, 78.8%, and 54.5% respectively. Panels a, b, c, and f within Figures 4.27 and 4.28 show that western and northern-peninsular India experienced more severe and widespread droughts during the period 1902-1935 as compared to other regions. This again highlights the differences in drought characteristics between regions. Regional accounting of drought intensity and areal extent is therefore useful for decision-makers in determining the impact of past and current droughts over each region and taking both short-term and long-term policy decisions.

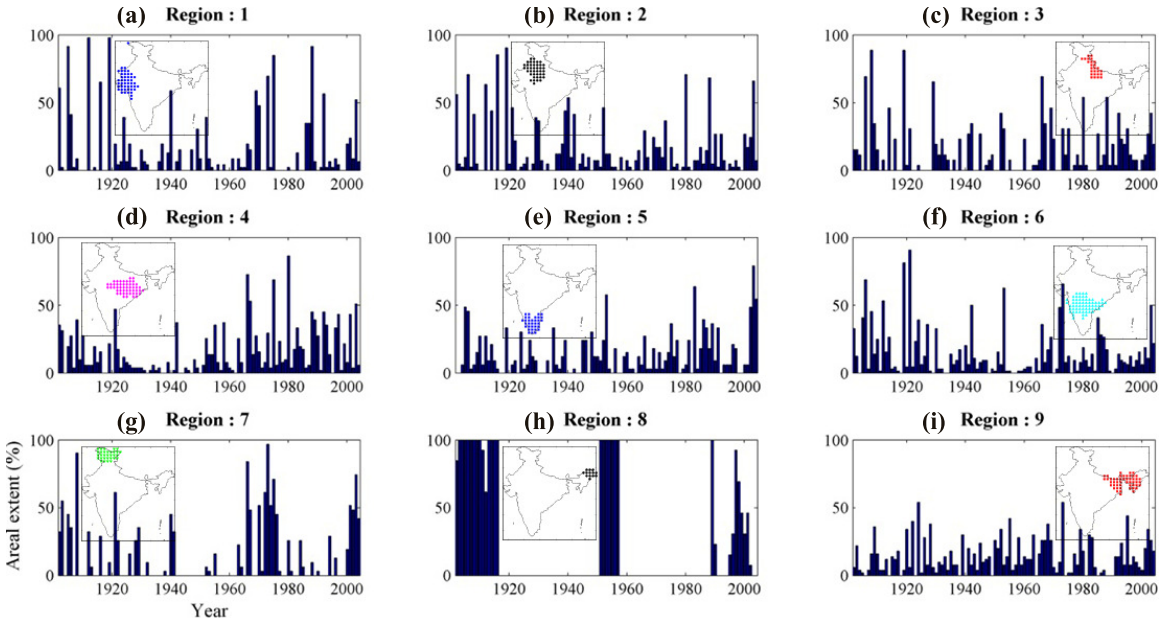

Figure 4.28. Areal extent of SPI-12 droughts ending in May for 1901-2004 over nine homogeneous regions (a) - (i) over India. SPI-12 intensities were computed after pooling precipitation data over homogeneous drought regions. The geographic extent of each homogeneous region is shown as an inset within the subplots.

Figure 4.29 shows bar plots of SPI-4 intensities for September ending months during the period 1902 to 2004 at nine SPI-4 homogeneous drought regions over India. The geographic extent of each region is shown as an inset within the sub-plots. This figure is useful for identifying major drought events of a 4-month time window across different regions in India. For example, region 1 or the western coast of India (Figure 4.29a) experienced extreme droughts during the years 1918, 1987, and 2002 with SPI-4 intensities of -3.46, -2.27, and -2.64, respectively. Similarly, the neighbouring region 2 or peninsular India (Figure 4.29b) experienced extreme droughts during the years 1918, 1920, 1952, and 1972 with SPI-4 intensities of -2.97, -2.60, -2.0, and -2.08, respectively. In contrast to region 1, it experienced moderate droughts during 1987 and 2002 of SPI-4 intensities -1.0 and -1.67, respectively.

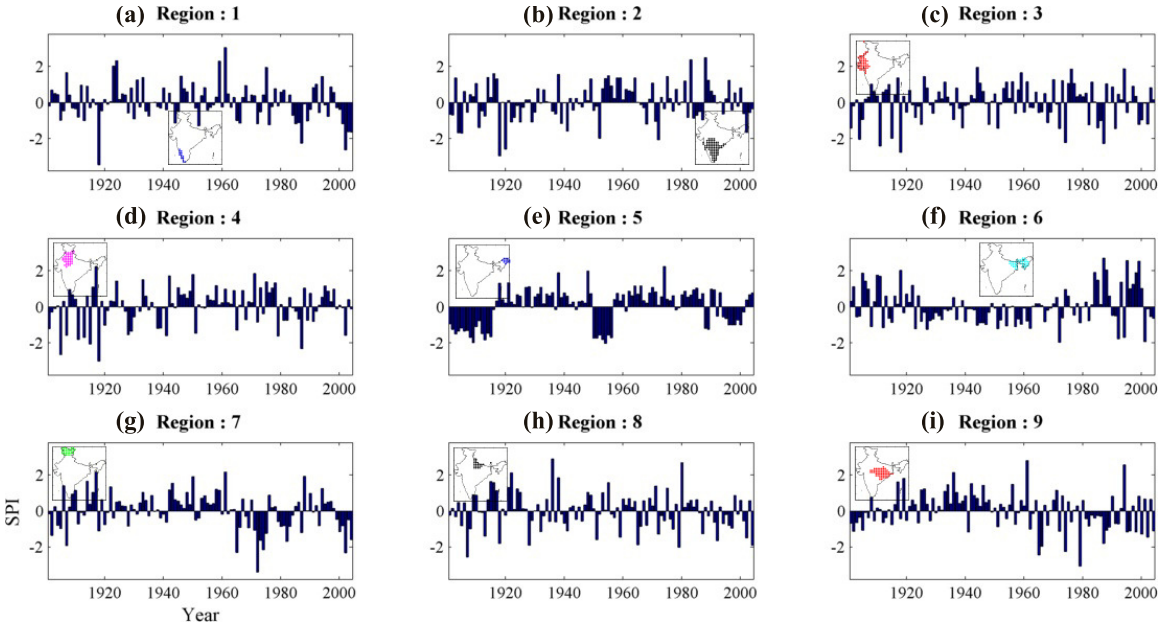

Figure 4.29. SPI-4 drought intensity ending in September for 1901-2004 computed after pooling precipitation data over nine homogeneous drought regions (a) - (i) over India. The geographic extent of each homogeneous region is shown as an inset within the subplots.

Figure 4.30 shows bar plots of the areal extent (in percentage) of SPI-4 droughts ending in September at each of the nine SPI-4 homogeneous drought regions over

India during the period 1902 to 2004. The geographic extent of each region is shown as an inset within each sub-plot. Figure 4.30a shows that the areal extent of SPI-4 droughts of extreme category in region 1 or western coast of India was 92.9%, 78.6%, and 85.7% during the September months of 1918, 1987, and 2002, respectively. Panels c, d, and h within Figures 4.29 and 4.30 show that western India and northern parts of the Indo-Gangetic plains experienced more number of severe to extreme, widespread droughts during the period 1902-1935 compared to other regions.

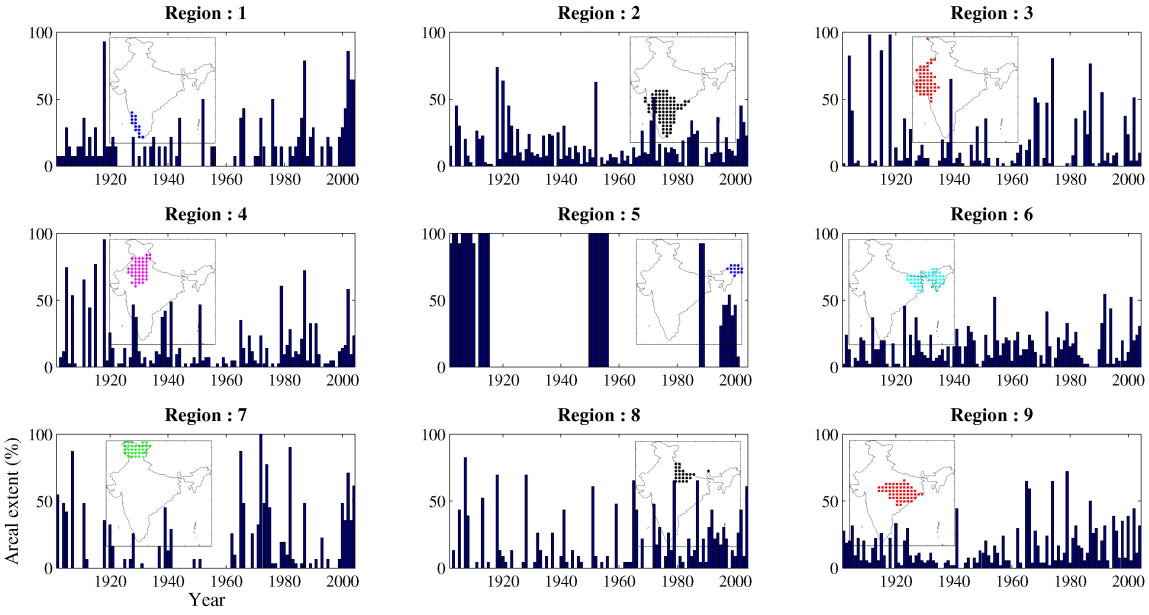

Figure 4.30. Areal extent of SPI-4 droughts ending in September for 1901-2004 over nine homogeneous regions (a) - (i) over India. SPI-4 computation was carried out after pooling precipitation data over homogeneous drought regions. The geographic extent of each homogeneous region is shown as an inset within the subplots.

The vulnerability of a region to droughts depends on several factors such as the total population (i.e., maintaining a reliable supply of drinking water), land use activities (i.e., agriculture), presence of critical infrastructure (e.g., power plants, reservoirs, etc.), inland navigation needs, industries (e.g., explicit or implicit need of water), etc. Figure 4.31 shows bar charts of total population and percentage of land use belonging

to six broad classes within each homogenous drought cluster. The geographic extent of each region is shown as an inset within each sub-plot [labeled (a) - (i) in Figure 4.28, instead of (1) - (9)] of Figure 4.28. According to Figure 4.31a, Region 9 belonging to parts of West Bengal, Jharkhand, Bihar, and northeastern states of Assam, Meghalaya, Tripura, and Mizoram - is the most populated with a total population greater than 350 million. Similarly, Region 7 and 8, belonging to Jammu & Kashmir and parts of some northeastern states such as Assam, Arunachal Pradesh, Nagaland, and Manipur, respectively, are sparsely populated (approximately 50 million people) compared to rest of India. Each of the remaining six homogeneous drought regions has a population greater than 140 million. Figure 4.31b shows the percentage of land use class within each homogeneous region. The six broad classes are listed in the caption of Figure 4.31. Agriculture is the dominant land use (greater than 40%) in all regions except regions 7 and 8. Forests are a dominant land cover in region 8 (64%), with significant coverage also in regions 9 (34%) and 4 (32%). Grasslands are dominant in regions 7 (21%) and 1 (11%), respectively. Snow and ice caps are dominant in region 3 (9 %), region 7 (31%), and region 8 (13%), respectively. By combining drought characteristics, total population, and land use distribution within each region, decision-makers may be able to identify most vulnerable regions to droughts (Schwarz et al., 2020). For example, regions 2, 6, and 9 are the most vulnerable regions based on population and agricultural land use information.

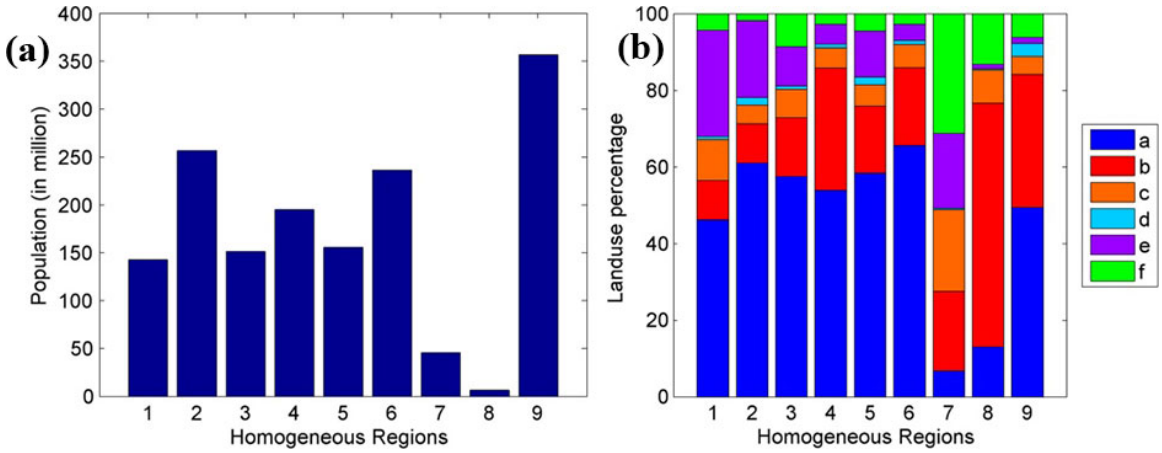

Figure 4.31. Identifying regions over India that are vulnerable to droughts using (a) total population and (b) broad land use and land cover classes. Legend description for land use and land cover classes: (a) Agricultural, (b) Forests, (c) Grassland, (d) Urban, (e) Barren, and (f) Others (e.g., water bodies, ice caps and snow, salt pans, etc.)

4.5 Summary and concluding remarks

In this study, homogeneous drought regions over India were identified using (i) multiple clustering methods, (ii) multiple drought indices, and (iii) for two time-windows of analysis. Three clustering algorithms - k-means, graph cuts, and agglomerative technique - that differ in their mathematical formulation were used. The following drought indices (i) SPI, (ii) SPEI, (iii) probabilistic SPI (pSPI), and (iv) combination of the three drought indices were used to describe drought characteristics at different grids over India. In addition to drought intensity values, geographic information of the grids in the form of latitude and longitude were used. Land use and land cover data were also used as inputs to the clustering algorithms. The Davis-Bouldin index was used to evaluate the compactness and the optimal number of clusters. Both short-term (4-month time window) and long term droughts (12-month time window) were considered when identifying homogeneous drought regions.

The study found that each algorithm produced different numbers of clusters, and their geographic extent varied significantly based on the choice of drought index. It was also observed that for any drought index, the homogeneous drought clusters differed depending on the choice of the clustering algorithm. The study also revealed that clustering results were sensitive to the length of the input dataset and was tested by generating results for different study periods - (i) 1901 to 1935, (ii) 1935 to 1970, (iii) 1971 to 2004, and (iv) 1901 to 2004. The sensitivity to the dimensionality of input datasets was presented for the following predictor sets (i) drought index (e.g., SPI) (ii) drought index and geographical information combined, and (iii) drought index, geographic information, and land use data. The clustering results were also sensitive to the choice of model parameters (such as energy constant value in graph cuts algorithm, initialization technique in k-means, or selection of linkage criteria used in the agglomerative method), cluster evaluation metric, spatial resolution of inputs (1°, 0.25° grids, etc.), and the spatial extent of the study domain (core-monsoon region). Among the three clustering algorithms, graph cuts provide a framework of implicitly imposing geographic contiguity through its MRF prior. When geographic information are provided as inputs, geographic contiguity is imposed explicitly. However, when high spatial resolution data were used (e.g., 0.25° grids), geographic information turned out to be a weak input for deciding cluster assignments.

Next, an application of clustering of clusters using two different methodologies: (i) Similarity matrix and (ii) Connected-triple-based similarity were presented for the first time in the case of homogeneous drought regionalization. A framework for combining clustering results is particularly useful when the resulting clusters are sensitive to user inputs, as shown in this study. Geographic contiguity is not imposed explicitly during the clustering of clusters. However, they may be implicitly imposed through the choice of base clustering algorithms (e.g., graph cuts) or input datasets used during base clustering (e.g., geographic data such as latitude, longitude, etc.). The clustering of clusters is not guaranteed to give globally optimal solution, as it depends on whether the individual base clustering methods cover the whole space

of solutions (i.e., the completeness assumption). Unlike supervised classification or clustering methods where ground truth labels may be used to compute homogeneity and completeness statistics (Rosenberg and Hirschberg, 2007), such an evaluation is not possible for unsupervised clustering methods. The clustering of clusters is similar to an ensemble model where outputs from individual base models are weighted to produce the final output. The accuracy of clustering of clusters is dependent on how well each base clustering model performs and how different each base models are in their modeling approach (Ghosh and Acharya, 2014; Kuncheva and Hadjitodorov, 2004).

The final regions obtained from an ensemble of base clusters were then used to analyze regional drought characteristics. Intensity-duration-frequency curves and Intensity-area-frequency curves were developed for each homogeneous area that allows water resources planners to assess the return periods of current and past droughts using features such as intensity, duration, and areal extent. Other regional drought characteristics such as average intensity during drought events, the length of drought events, and the number of drought events at each grid over homogeneous regions were presented in the form of box plots. Such box plots are useful in highlighting the general distribution of drought characteristics over an area. Regionalization also provides a framework for pooling data over for drought or other hydrometeorological analysis. In this study, a representative precipitation time series was first obtained by pooling precipitation data at all grids within a region. The representative precipitation data was later used to identify significant historic droughts and their characteristics such as intensity and areal extent over each homogeneous drought region in India and assess how these have evolved over the study period.

To conclude, methods to identify homogeneous drought regions at a national scale (over India) were presented using multiple clustering methods and datasets. The study found that homogeneous drought regions were sensitive to the choice of (i) clustering algorithm (k-means, graph cuts, hierarchical, etc.), (ii) drought index (SPI, SPEI, etc.), (iii) time windows (SPI-4, SPI-12, etc.), (iv) period of analysis (1936-1970

vs. 1971-2004, etc.), (v) predictor set (drought characteristics, land use, geographic information, etc.), (vi) input data resolution (1° versus 0.25° grids), and (vii) spatial domain of analysis (e.g. core-monsoon region). It is therefore recommended that homogeneous drought regions should be analyzed from both short-term (4-month window) and long-term (12-month window) planning objective. For each objective, homogeneous regions should be developed using multiple drought indices and base clustering methods. Then using either pair-wise similarity matrix or connected triple-based similarity matrix approach the unique clusters obtained from base clustering methods may be combined to get final clusters. The final drought clusters denoting areas with similar drought characteristics are expected to help policy-makers and water resources planners in the optimal allocation of resources, developing drought management plans and taking timely actions to mitigate the negative impacts during droughts.

The methodology presented in this study captures homogeneity mainly in terms of drought intensities, duration of drought events, and timing of droughts (i.e., onset and termination). Geographic information and land use data also play a role in final cluster formation. While the quality of homogeneity is difficult to assess in the absence of ground truth labels, the drought characteristics within each region were compared in the study using box plots for this purpose. The validation of homogeneity would provide confidence to combine data over an area for other studies, e.g., multi-year drought analysis. The clustering results presented in this study show discrete cluster assignments. However, clustering algorithms such as graph cuts are capable of providing probabilistic cluster assignments, similar to those achieved using the fuzzy c-means algorithm. The knowledge about uncertainty in cluster assignment can be useful during (a) policy making (e.g., demarcating regional boundaries, etc.), (b) pooling data for regional analysis (i.e., exclusion of grids), etc. The motivation for conducting regionalization at a larger regional or national scale is that droughts themselves may extend across several state lines. Though significant policy decisions and allocation of resources are made at the regional level, the authority to

implement drought management and mitigation measures may lie with the states and local authorities (MacDonald, 2007; Wilhite, 2012). Drought response actions may thus require decision-makers from different states, counties, villages, etc., to evaluate what would be a good plan for each of their jurisdictions. Regional drought characteristics can also help in designing local and regional infrastructure design. For example, drought characteristics over a region can be used in the sizing and operation of reservoirs based on the probability of occurrence of a *design* drought (Hudson and Roberts, 1955; Shih and ReVelle, 1994), farm ponds (Panigrahi and Panda, 2003), and maintaining stream ecosystem services (Atkinson et al., 2014).

Regionalization studies, when used in conjunction with crop models, can help in estimating the impact of droughts on crop yield. Crop models use weather data (precipitation, soil moisture, temperature, sunlight, etc.), soil data, and crop management data to simulate the agricultural plant growth patterns over a study domain, such as field, village, or a county. The output from crop models provides useful insight into the relationship between drought characteristics - such as intensity, duration, and timing during a growing season, and reduced crop yield. This knowledge can then be used to estimate crop yields over a broader region with similar drought characteristics. Policy-makers can use this information to prescribe best farming practices over the area during drought events, and shape watershed and regional policy development (Motha, 2011). If drought forecasts are available at the regional scale, they may be used as inputs to the crop models to obtain modified irrigation schedules to compensate moisture deficits depending on the crop and its growth stage. Long-term regional drought characteristics can be used as inputs to crop models to estimate the risk in reduction of crop yield over the region for different drought scenarios, e.g., severe drought (Leng and Hall, 2019).

The following limitations of this study may be addressed as part of future work:

1. All input datasets such as precipitation time series, drought characteristics, and land use data were assumed to be stationary irrespective of the chosen period of analysis (35 year epochs or entire study period). The climate community has

generally accepted 30-year periods as the length of time over which stationary assumption can be made (Arguez et al., 2012). A more rigorous approach would be to conduct a change-point analysis to identify periods over which stationary assumptions are valid (Aminikhanghahi and Cook, 2017), and then obtain homogeneous clusters over these periods. However, this can be challenging when considering multiple datasets because each dataset may have different change points.

2. Only 4-month and 12-month drought time windows were considered for identifying homogeneous drought regions. When considering multi-year droughts, if overlapping datasets are used then the assumption of independence will be violated. On the other hand, the data available for clustering would be thinned down significantly when non-overlapping inputs are used.

3. Additional datasets such as soil moisture and streamflow data may be used for clustering.

4. Only a limited number of base clustering algorithms and drought indices were used in the study.

5. SUMMARY

Droughts are natural disasters caused due to water deficit expressed through hydrometeorological variables such as precipitation, soil moisture, streamflow, etc. It is often challenging to characterize droughts promptly as they manifest gradually and are already a daunting event by the time their presence is recognized. Several drought indices such as SPI, SPEI, etc. are used to monitor drought characteristics such as intensity and duration. These indices are also used to classify droughts into different categories, such as moderate, severe, and extreme. These drought classes are then used to trigger responses from decision-makers to mitigate the negative impacts of droughts. However, experts believe that the allocation of resources and response capabilities of communities will benefit from the use of drought indices that provide estimates of model uncertainty while classifying droughts. Characterization of droughts using probabilistic models that account for model uncertainties is presented in this thesis. The main findings of the study are summarized below.

1. Based on the recommendation by Trenberth et al. (2014), long-term retrospective drought variability was examined over the Indian Monsoon Region (IMR) using multiple datasets and methods. While some specific differences in results were observed based on the choice of datasets and methods, the overall conclusions were consistent. Results indicate droughts over IMR are becoming more regional in recent decades. The Indo-Gangetic plain, parts of coastal south-India, and central Maharashtra were identified as vulnerable regions for recent droughts.

2. A probabilistic Gamma mixture model-based drought index was presented as an alternative to (a) deterministic classification by SPI, and (b) probabilistic classification by HMM-DI. The Bayesian framework of the proposed model avoids

over-specification and overfitting by choosing the optimum number of mixture components required to model the data − a problem that is often encountered in other probabilistic drought indices (e.g., HMM-DI). When a sufficient number of components are used in Gamma-MM, it can provide a good approximation to any continuous distribution in the range $(0,\infty)$, thus addressing the problem of choosing an appropriate distribution for SPI analysis. The Gamma-MM propagates model uncertainties to drought classification.

3. Finding regions over a study area that have similar drought characteristics is useful for drought preparedness and management purposes, resource allocation, thereby improving the overall resilience of different regions to droughts. The regionalization framework proposed in this study identified grids that exhibit homogeneity with respect to drought intensity values, duration of drought events, and their timing (onset and termination of drought events). Drought characteristics such as intensity, frequency, and duration, along with land-use and geographic information, were used as input features for clustering algorithms. Three methods, namely, (i) a Bayesian graph cuts algorithm that combines the Gaussian mixture model (GMM) and Markov random fields (MRF), (ii) k-means, and (iii) hierarchical agglomerative clustering algorithm were used to find homogeneous drought regions that are spatially contiguous and have similar drought characteristics. The number of homogeneous clusters and their shape were found to be sensitive to the choice of the drought index, time window of drought, period of analysis, dimensionality of input datasets, clustering method, spatial resolution of input datasets, spatial extent of the study domain, and model parameters of clustering algorithms. Regionalization for different epochs provided useful insights into the space-time evolution of homogenous drought regions over the study area. Strategies to combine the results from multiple clustering methods were presented. The accuracy of clustering of clusters, which is similar to ensemble modeling or averaging in machine learning, depends on how each clustering model performs and the uniqueness in their

modeling approach. The combination of multiple weak models will likely lead to a robust model. However, since ground truth labels are unavailable, it is not possible to confidently determine if the clusters are valid or if the globally optimal solution was reached. The nature of drought characteristics, intensity-duration-frequency curves, and intensity-area-frequency curves were developed over each homogenous drought region. These results can help policy-makers and water resources planners in the optimal allocation of resources, developing drought management plans, and taking timely actions to mitigate the negative impacts during droughts.

Assessing the causal mechanism of droughts, and relating trends in drought characteristics to phenomena such as changes observed in the monsoon break (active - dry spell) periods (Singh et al., 2014), aerosols, land use, SST, thermodynamic feedback due to heating rates (Roxy et al., 2015) were not considered in this study. Also, the precipitation and temperature time series, drought characteristics, and land use data were assumed to be stationary. The results presented in this study were only for 4-month, 7-month, and 12-month drought time windows - while some users may be interested in drought characteristics for shorter (1-month, 2-month, etc.) or longer time windows (24-month, 36-month, etc.). Data available for analyzing longer time window droughts is significantly limited when analysis is performed on a station-by-station basis. However, as long term droughts have large spatial extents, identifying homogeneous regions over a study area allows users to pool the data from multiple stations within a region that have similar drought characteristics and conduct robust studies. The spatial evolution of drought characteristics and clusters were analyzed for only three 35-year epochs. While the choice of epoch-length is subjective (chosen here based on climate normals, i.e., 30 years or more), stationarity was assumed to be valid over this time-frame. Alternatively, methods such as change-point detection may be used to select the length of epochs. Droughts due to deficit in water expressed through other hydrometeorological variables such as soil moisture and streamflow data

were not considered. The homogeneous drought clusters obtained in this study could not be validated with ground truth labels - as these do not exist.

Future work will address some of the above limitations. Probabilistic drought indices that account for non-stationarity and auto-correlation in hydrometeorological data will be developed. Methods to relate changes observed in drought characteristics over India to anthropogenic factors such as greenhouse gas emissions and aerosol concentrations, land use changes, etc. will be explored. Future studies should also consider including soil moisture and streamflow deficits for characterizing agricultural and hydrologic droughts. Regionalization methods that account for non-stationarity in hydrometeorological and land use data, engage stakeholder feedback and features such as administrative boundaries will be explored.

REFERENCES

Abatzoglou, J. T., Redmond, K. T., and Edwards, L. M. (2009). Classification of regional climate variability in the state of California. *Journal of Applied Meteorology and Climatology*, 48(8):1527–1541.

AghaKouchak, A. (2014). A baseline probabilistic drought forecasting framework using standardized soil moisture index: application to the 2012 United States drought. *Hydrology and Earth System Sciences*, 18(7):2485–2492.

Ahmed, B. Y. M. (1997). Climatic classification of Saudi Arabia: An application of factor–cluster analysis. *GeoJournal*, 41(1):69–84.

Aldrian, E. and Dwi Susanto, R. (2003). Identification of three dominant rainfall regions within Indonesia and their relationship to sea surface temperature. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 23(12):1435–1452.

Alijani, B., Ghohroudi, M., and Arabi, N. (2008). Developing a climate model for Iran using GIS. *Theoretical and Applied Climatology*, 92(1-2):103–112.

Allen, R. G., Pereira, L. S., Raes, D., Smith, M., and others (1998). Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. *FAO, Rome*, 300(9):D05109.

Alvarez, J. and Estrela, T. (2003). Regionalisation and identification of droughts in mediterranean countries of Europe. In *Tools for drought mitigation in Mediterranean Regions*, pages 123–146. Springer.

Aminikhanghahi, S. and Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367.

Arguez, A., Durre, I., Applequist, S., Vose, R. S., Squires, M. F., Yin, X., Heim Jr, R. R., and Owen, T. W. (2012). NOAA's 1981–2010 US climate normals: An overview. *Bulletin of the American Meteorological Society*, 93(11):1687–1697.

Atkinson, C. L., Julian, J. P., and Vaughn, C. C. (2014). Species and function lost: role of drought in structuring stream communities. *Biological Conservation*, 176:30–38.

Ayad, H. and Kamel, M. (2003). Finding natural clusters using multi-clusterer combiner based on shared nearest neighbors. In *International Workshop on Multiple Classifier Systems*, pages 166–175. Springer.

Bagla, P. (2006). Controversial rivers project aims to turn India's fierce monsoon into a friend. *Science*, 313(5790):1036–1037.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300. ArticleType: research-article / Full publication date: 1995 / Copyright 1995 Royal Statistical Society.

Bharath, R. and Srinivas, V. (2015). Delineation of homogeneous hydrometeorological regions using wavelet-based global fuzzy cluster analysis. *International Journal of Climatology*, 35(15):4707–4727.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, volume 1. Springer New York.

Bonaccorso, B., Peres, D. J., Cancelliere, A., and Rossi, G. (2013). Large scale probabilistic drought characterization over Europe. *Water Resources Management*, 27(6):1675–1692.

Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11):1222–1239.

Bravar, L. and Kavvas, M. (1991). On the physics of droughts. ii. Analysis and simulation of the interaction of atmospheric and hydrologic processes during droughts. *Journal of Hydrology*, 129(1-4):299–330.

Burn, D. H. and Elnur, M. A. H. (2002). Detection of hydrologic trends and variability. *Journal of Hydrology*, 255(1-4):107–122.

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

Center for International Earth Science Information Network - CIESIN - Columbia University, United Nations Food and Agriculture Programme - FAO, and Centro Internacional de Agricultura Tropical - CIAT (2005). Gridded Population of the World, Version 3 (GPWv3): Centroids.

Charusombat, U. and Niyogi, D. (2011). A Hydroclimatological Assessment of Regional Drought Vulnerability: A Case Study of Indiana Droughts. *Earth Interactions*, 15(26):1–65.

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

Chung, C. E. and Ramanathan, V. (2006). Weakening of North Indian SST Gradients and the Monsoon Rainfall in India and the Sahel. *Journal of Climate*, 19(10):2036–2045.

Clifford, P. (1990). Markov random fields in statistics. *Disorder in physical systems: A volume in honour of John M. Hammersley*, 19.

Cole, J. E. and Cook, E. R. (1998). The changing relationship between ENSO variability and moisture balance in the continental United States. *Geophysical Research Letters*, 25(24):4529–4532.

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

Dai, A. (2013). Increasing drought under global warming in observations and models. *Nature Climate Change*, 3(1):52–58.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

De, U., Dube, R., and Rao, G. P. (2005). Extreme weather events over India in the last 100 years. *J. Ind. Geophys. Union*, 9(3):173–187.

Degaetano, A. T. (1996). Delineation of mesoscale climate zones in the northeastern United States using a novel approach to cluster analysis. *Journal of Climate*, 9(8):1765–1782.

DeGaetano, A. T. (2001). Spatial grouping of United States climate stations using a hybrid clustering approach. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 21(7):791–807.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38. ArticleType: research-article / Full publication date: 1977 / Copyright 1977 Royal Statistical Society.

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive approximation*, volume 303. Springer.

Dracup, J. A., Lee, K. S., and Paulson Jr, E. G. (1980). On the definition of droughts. *Water resources research*, 16(2):297–302.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Evin, G., Merleau, J., and Perreault, L. (2011). Two-component mixtures of normal, gamma, and Gumbel distributions for hydrological applications (W08525). *Water Resources Research*, 47(8).

Federal Emergency Management Agency (1995). National mitigation strategy: Partnerships for building safer communities. *Mitigation directorate*.

Fred, A. L. and Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. *IEEE transactions on pattern analysis and machine intelligence*, 27(6):835–850.

Gadgil, S. and Joshi, N. (1983). Climatic clusters of the Indian region. *Journal of Climatology*, 3(1):47–63.

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

Ghosh, J. and Acharya, A. (2014). *Cluster Ensembles: Theory and Applications*. CRC Press, Boca Raton FL.

Ghosh, S. and Srinivasan, K. (2016). Analysis of spatio-temporal characteristics and regional frequency of droughts in the southern peninsula of India. *Water resources management*, 30(11):3879–3898.

Glantz, M. H. and Orlovsky, N. S. (1983). Desertification: A review of the concept. *Desertification Control Bulletin*, 9:15–22.

Goswami, B. N., Venugopal, V., Sengupta, D., Madhusoodanan, M. S., and Xavier, P. K. (2006). Increasing trend of extreme rain events over India in a warming environment. *Science*, 314(5804):1442 –1445.

Gowda, K. C. and Krishna, G. (1978). Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern recognition*, 10(2):105–112.

Goyal, M. K. and Sharma, A. (2016). A fuzzy c-means approach regionalization for analysis of meteorological drought homogeneous regions in western India. *Natural Hazards*, 84(3):1831–1847.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Guhathakurta, P. and Rajeevan, M. (2008). Trends in the rainfall pattern over India. *International Journal of Climatology*, 28(11):1453–1469.

Guttman, N. B. (1998). Comparing the palmer drought index and the standardized precipitation index. *JAWRA Journal of the American Water Resources Association*, 34(1):113–121.

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

Hamed, K. H. and Ramachandra Rao, A. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, 204(14):182–196.

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

Hayes, M. J., Wilhelmi, O. V., and Knutson, C. L. (2004). Reducing drought risk: bridging theory and practice. *Natural Hazards Review*, 5(2):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):1149.

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

Houghton, J. T., Ding, Y., Griggs, D. J., Noguer, M., Linden, P. J. v. d., Dai, X., Maskell, K., and Johnson, C. A., editors (2001). *Climate change 2001: The scientific basis*. Cambridge University Press.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Hudson, H. E. and Roberts, W. J. (1955). 1952-1955 Illinois drought with special reference to impounding reservoir design. *Bulletin (Illinois State Water Survey) no. 43*.

Iam-on, N., Garrett, S., et al. (2010). Linkclue: A matlab package for link-based cluster ensembles. *Journal of Statistical Software*, 36(9):1–36.

Iyigun, C., Türkeş, M., Batmaz, İ., Yozgatligil, C., Purutçuoğlu, V., Koç, E. K., and Öztürk, M. Z. (2013). Clustering current climate regions of Turkey by using a multivariate statistical method. *Theoretical and applied climatology*, 114(1-2):95–106.

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

Karl, T. and Koss, W. J. (1984). Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983.

Kishtawal, C. M., Niyogi, D., Tewari, M., Pielke Sr, R. A., and Shepherd, J. M. (2010). Urbanization signature in the observed heavy rainfall climatology over India. *International Journal of Climatology*, 30(13):1908–1916.

Klink, S., Reuther, P., Weber, A., Walter, B., and Ley, M. (2006). Analysing social networks within bibliographical data. In *International Conference on Database and Expert Systems Applications*, pages 234–243. Springer.

Kothawale, D. and Rajeevan, M. (2017). Monthly, seasonal and annual rainfall time series for all-India, homogeneous regions and meteorological subdivisions: 1871–2016. *Contribution from IITM Research Report No. RR-138*.

Kripalani, R. H., Kulkarni, A., Sabade, S. S., and Khandekar, M. L. (2003). Indian Monsoon Variability in a Global Warming Scenario. *Natural Hazards*, 29(2):189–206. 10.1023/A:1023695326825.

Krishnamurthy, V. and Shukla, J. (2000). Intraseasonal and interannual variability of rainfall over India. *Journal of Climate*, 13(24):4366–4377.

Kulkarni, A. and von Storch, H. (1995). Monte Carlo experiments on the effect of serial correlation on the Mann-Kendall test of trend. *Meteorologische Zeitschrift*, 4(2):82–85.

Kumar, K. R., Pant, G. B., Parthasarathy, B., and Sontakke, N. A. (1992). Spatial and subseasonal patterns of the longterm trends of Indian summer monsoon rainfall. *International Journal of Climatology*, 12(3):257–268.

Kuncheva, L. I. and Hadjitodorov, S. T. (2004). Using diversity in cluster ensembles. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, volume 2, pages 1214–1219. IEEE.

Kuncheva, L. I. and Vetrov, D. P. (2006). Evaluation of stability of k-means cluster ensembles with respect to random initialization. *IEEE transactions on pattern analysis and machine intelligence*, 28(11):1798–1808.

Kyrgyzov, I. O., Maitre, H., and Campedel, M. (2007). A method of clustering combination applied to satellite image analysis. In *14th International Conference on Image Analysis and Processing (ICIAP 2007)*, pages 81–86. IEEE.

Kyselỳ, J., Picek, J., and Huth, R. (2007). Formation of homogeneous regions for regional frequency analysis of extreme precipitation events in the Czech Republic. *Studia Geophysica et Geodaetica*, 51(2):327–344.

Lana, X., Serra, C., and Burgueño, A. (2001). Patterns of monthly rainfall shortage and excess in terms of the standardized precipitation index for Catalonia (NE Spain). *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 21(13):1669–1691.

Law, M. H., Topchy, A. P., and Jain, A. K. (2004). Multiobjective data clustering. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE.

Leber, D., Holawe, F., and Häusler, H. (1995). Climatic classification of the Tibet Autonomous Region using multivariate statistical methods. *GeoJournal*, 37(4):451–472.

Leng, G. and Hall, J. (2019). Crop yield sensitivity of global major agricultural countries to droughts and the projected changes in the future. *Science of the Total Environment*, 654:811–821.

Li, Z. and Zhang, Y. (2011). Application of Gaussian Mixture Model and Estimator to Radar-Based Weather Parameter Estimations. *Geoscience and Remote Sensing Letters, IEEE*, 8(6):1041 –1045.

Liang, Z., Wang, D., Guo, Y., Zhang, Y., and Dai, R. (2011). Application of Bayesian Model Averaging Approach to MultiModel Ensemble Hydrologic Forecasting. *Journal of Hydrologic Engineering*, 1(1):326–326.

Liu, W. T. and Jurez, R. I. N. (2001). ENSO drought onset prediction in northeast Brazil using NDVI. *International Journal of Remote Sensing*, 22(17):3483–3501.

Lloyd-Huges, B. and Saunders, M. A. (2002). A drought climatology for Europe. *International Journal of Climatology*, 22:1571–1592.

Loukas, A. and Vasiliades, L. (2004). Probabilistic analysis of drought spatiotemporal characteristics in Thessaly region, Greece. *Natural Hazards and Earth System Science*, 4(5/6):719–731.

MacDonald, G. M. (2007). Severe and sustained drought in southern California and the West: Present conditions and insights from the past on causes and impacts. *Quaternary International*, 173:87–100.

Madadgar, S. and Moradkhani, H. (2013). A Bayesian Framework for Probabilistic Seasonal Drought Forecasting. *Journal of Hydrometeorology*, 14(6):1685–1705.

Mallya, G. (2011). Hidden Markov model based probabilistic assessment of droughts. Master's thesis, Purdue University, West Lafayette, IN 47907.

Mallya, G., Mishra, V., Niyogi, D., Tripathi, S., and Govindaraju, R. S. (2016). Trends and variability of droughts over the Indian monsoon region. *Weather and Climate Extremes*, 12:43–68.

Mallya, G., Tripathi, S., and Govindaraju, R. S. (2015). Probabilistic drought classification using gamma mixture models. *Journal of Hydrology*, 526:116–126.

Mallya, G., Tripathi, S., Kirshner, S., and Govindaraju, R. S. (2013). Probabilistic Assessment of Drought Characteristics Using Hidden Markov Model. *Journal of Hydrologic Engineering*, 18(7):834–845.

Malmgren, B. A. and Winter, A. (1999). Climate zonation in Puerto Rico based on principal components analysis and an artificial neural network. *Journal of climate*, 12(4):977–985.

Matulla, C., Penlap, E. K., Haas, P., and Formayer, H. (2003). Comparative analysis of spatial and seasonal variability: Austrian precipitation during the 20th century. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 23(13):1577–1588.

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

McKee, T. B., Doesken, N. J., and Kleist, J. (1995). Drought monitoring with multiple time scales. In *Proceedings of the 9th Conference on Applied Climatology*, pages 233–236. American Meteorological Society Dallas, Boston, MA.

McLachlan, G. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* John Wiley & Sons, Inc., New York.

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

Mishra, A. K. and Singh, V. P. (2010). A review of drought concepts. *Journal of Hydrology*, 391(12):202–216.

Mishra, A. K., Singh, V. P., and Desai, V. R. (2009). Drought characterization: a probabilistic approach. *Stochastic Environmental Research and Risk Assessment*, 23(1):41–55.

Mishra, V., Smoliak, B. V., Lettenmaier, D. P., and Wallace, J. M. (2012). A prominent pattern of year-to-year variability in Indian Summer Monsoon Rainfall. *Proceedings of the National Academy of Sciences*.

Mo, K. C. (2008). Model-Based Drought Indices over the United States. *Journal of Hydrometeorology*, 9:1212–1230.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2):91–118.

Mooley, D., Parthasarathy, B., Sontakke, N., and Munot, A. (1981). Annual rainwater over India, its variability and impact on the economy. *Journal of Climatology*, 1(2):167–186.

Motha, R. P. (2011). Use of crop models for drought analysis. *Drought Mitigation Center Faculty Publications*, 58.

Nguyen, N. and Caruana, R. (2007). Consensus clusterings. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 607–612. IEEE.

Niranjan Kumar, K., Rajeevan, M., Pai, D. S., Srivastava, A. K., and Preethi, B. (2013). On the observed variability of monsoon droughts over India. *Weather and Climate Extremes*, 1:42–50.

Niyogi, D., Kishtawal, C., Tripathi, S., and Govindaraju, R. S. (2010). Observational evidence that agricultural intensification and land use change may be reducing the Indian summer monsoon rainfall. *Water Resources Research*, 46(3):n/a–n/a.

Pai, D., Sridhar, L., Rajeevan, M., Sreejith, O., Satbhai, N., and Mukhopadhyay, B. (2014). Development of a new high spatial resolution (0.25× 0.25) long period (1901–2010) daily gridded rainfall data set over india and its comparison with existing data sets over the region. *Mausam*, 65(1):1–18.

Palmer, W. C. (1965). *Meteorological drought*, volume 30. US Department of Commerce, Weather Bureau.

Pan, M., Yuan, X., and Wood, E. F. (2013). A probabilistic framework for assessing drought recovery. *Geophysical Research Letters*, 40(14):3637–3642.

Panigrahi, B. and Panda, S. N. (2003). Optimal sizing of on-farm reservoirs for supplemental irrigation. *Journal of irrigation and drainage engineering*, 129(2):117–128.

Pappenberger, F. and Beven, K. J. (2006). Ignorance is bliss: Or seven reasons not to use uncertainty analysis. *Water Resources Research*, 42(5):W05302.

Parathasarathy, B., Munot, A., and Kothawale, D. (1994). Droughts over homogeneous regions of India: 18711990. *Drought Network News (1994-2001)*, page 67.

Parthasarathy, B., Kumar, K. R., and Munot, A. (1993). Homogeneous Indian monsoon rainfall: variability and prediction. *Proceedings of the Indian Academy of Sciences-Earth and Planetary Sciences*, 102(1):121–155.

Penman, H. L. (1948). Natural evaporation from open water, bare soil and grass. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 193, pages 120–145. The Royal Society.

Priestley, C. and Taylor, R. (1972). On the assessment of surface heat flux and evaporation using large-scale parameters. *Monthly weather review*, 100(2):81–92.

Puvaneswaran, M. (1990). Climatic classification for Queensland using multivariate statistical techniques. *International Journal of Climatology*, 10(6):591–608.

Rajeevan, M. (2006). High resolution daily gridded rainfall data for the Indian region: Analysis of break and active monsoon spells. *Current Science*, 91(3):296.

Rajeevan, M., Bhate, J., and Jaswal, A. K. (2008). Analysis of variability and trends of extreme rainfall events over India using 104 years of gridded daily rainfall data. *Geophysical Research Letters*, 35(18):L18707.

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

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420.

Rossi, G. and Cancelliere, A. (2003). At-site and regional drought identification by Redim model. In Rossi, G., Cancelliere, A., Pereira, L. S., Oweis, T., Shatanawi, M., and Zairi, A., editors, *Tools for drought mitigation in Mediterranean regions*, number 44 in Water Science and Technology Library, pages 37–54. Springer Netherlands.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Roxy, M. K., Ritika, K., Terray, P., Murtugudde, R., Ashok, K., and Goswami, B. (2015). Drying of Indian subcontinent by rapid Indian Ocean warming and a weakening land-sea thermal gradient. *Nature communications*, 6.

Roy, P., Meiyappan, P., Joshi, P., Kale, M., Srivastav, V., Srivasatava, S., Behera, M., Roy, A., Sharma, Y., Ramachandran, R., et al. (2016). Decadal land use and land cover classifications across India, 1985, 1995, 2005. *ORNL DAAC*.

Rupa Kumar, K., Sahai, A. K., Krishna Kumar, K., Patwardhan, S. K., Mishra, P. K., Revadekar, J. V., Kamala, K., and Pant, G. B. (2006). High-resolution climate change scenarios for India for the 21st century. *Current Science*, 90(3):334–345.

Russo, S., Dosio, A., Sterl, A., Barbosa, P., and Vogt, J. (2013). Projection of occurrence of extreme dry-wet years and seasons in Europe with stationary and nonstationary standardized precipitation indices. *Journal of Geophysical Research: Atmospheres*, 118(14):7628–7639.

Ryu, J. H., Svoboda, M. D., Lenters, J. D., Tadesse, T., and Knutson, C. L. (2010). Potential extents for ENSO-driven hydrologic drought forecasts in the United States. *Climatic Change*, 101(3-4):575–597.

Sahin, S. and Cigizoglu, H. K. (2012). The sub-climate regions and the sub-precipitation regime regions in Turkey. *Journal of hydrology*, 450:180–189.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Schwarz, M., Landmann, T., Cornish, N., Wetzel, K.-F., Siebert, S., and Franke, J. (2020). A spatially transferable drought hazard and drought risk modeling approach based on remote sensing data. *Remote Sensing*, 12(2):237.

Sheffield, J., Wood, E. F., and Roderick, M. L. (2012). Little change in global drought over the past 60 years. *Nature*, 491(7424):435–438.

Shiau, J.-T., Feng, S., and Nadarajah, S. (2007). Assessment of hydrological droughts for the Yellow River, China, using copulas. *Hydrological Processes*, 21(16):2157–2163.

Shih, J.-S. and ReVelle, C. (1994). Water-supply operations during drought: Continuous hedging rule. *Journal of Water Resources Planning and Management*, 120(5):613–629.

Singh, D., Tsiang, M., Rajaratnam, B., and Diffenbaugh, N. S. (2014). Observed changes in extreme wet and dry spells during the South Asian summer monsoon season. *Nature Climate Change*, 4(6):456–461.

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, Colorado1. *JAWRA Journal of the American Water Resources Association*, 41(1):11–24.

Srinivas, V., Tripathi, S., Rao, A. R., and Govindaraju, R. S. (2008). Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering. *Journal of Hydrology*, 348(1-2):148–166.

Stahl, K. and Demuth, S. (1999). Methods for regional classification of streamflow drought series: Cluster analysis. *ARIDE, Technical Rep*, 1.

Steinemann, A. (2003). *Drought indicators and triggers: A stochastic approach to evaluation*. Wiley Online Library.

Stephenson, D. (2001). Searching for a fingerprint of Global Warming in the Asian summer monsoon. *MAUSAM*, 52:213–213–220.

Stooksbury, D. and Michaels, P. (1991). Cluster analysis of southeastern US climate stations. *Theoretical and Applied Climatology*, 44(3-4):143–150.

Strehl, A. and Ghosh, J. (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617.

Svoboda, M., LeComte, D., Hayes, M., Heim, R., Gleason, K., Angel, J., Rippey, B., Tinker, R., Palecki, M., Stooksbury, D., et al. (2002). The drought monitor. *Bulletin of the American Meteorological Society*, 83(8):1181–1190.

Thornthwaite, C. W. (1948). An approach toward a rational classification of climate. *Geographical review*, pages 55–94.

Trenberth, K. E., Dai, A., van der Schrier, G., Jones, P. D., Barichivich, J., Briffa, K. R., and Sheffield, J. (2014). Global warming and changes in drought. *Nature Climate Change*, 4(1):17–22.

Tripathi, S. and Govindaraju, R. S. (2009). Change detection in rainfall and temperature patterns over India. In *Proceedings of the Third International Workshop on Knowledge Discovery from Sensor Data*, pages 133–141.

Trnka, M., Dubrovskỳ, M., Svoboda, M., Semerádová, D., Hayes, M., Žalud, Z., and Wilhite, D. (2009). Developing a regional drought climatology for the Czech Republic. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 29(6):863–883.

Unal, Y., Kindap, T., and Karaca, M. (2003). Redefining the climate zones of Turkey using cluster analysis. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 23(9):1045–1055.

Ventura, V., Paciorek, C. J., and Risbey, J. S. (2004). Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. *Journal of Climate*, 17(22):4343–4356.

Vicente-Serrano, S. M. (2006a). Differences in spatial patterns of drought on different time scales: an analysis of the Iberian peninsula. *Water resources management*, 20(1):37–60.

Vicente-Serrano, S. M. (2006b). Spatial and temporal analysis of droughts in the Iberian peninsula (1910–2000). *Hydrological Sciences Journal*, 51(1):83–97.

Vicente-Serrano, S. M., Beguera, S., and Lpez-Moreno, J. I. (2010). A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *Journal of Climate*, 23(7):1696–1718.

Wang, A., Lettenmaier, D. P., and Sheffield, J. (2011). Soil moisture drought in China, 1950–2006. *Journal of Climate*, 24(13):3257–3271.

Wilhite, D. A. (2000). Drought as a natural hazard: concepts and definitions. In Wilhite, D. A., editor, *Drought: A Global Assessment*, volume I, chapter 1, pages 3–18. London: Routledge.

Wilhite, D. A. (2012). *Drought assessment, management, and planning: theory and case studies: theory and case studies*, volume 2. Springer Science & Business Media.

Wilhite, D. A. and Svoboda, M. D. (2000). Drought early warning systems in the context of drought preparedness and mitigation. *Early warning systems for drought preparedness and drought management*, pages 1–21.

Wilks, D. S. (2006). On field significance and the false discovery rate. *Journal of Applied Meteorology and Climatology*, 45(9):1181–1189.

Wiper, M., Insua, D. R., and Ruggeri, F. (2001). Mixtures of gamma distributions with applications. *Journal of Computational and Graphical Statistics*, 10(3):440–454.

Yevjevich, V. M. (1967). An bjective approach to definitions and investigations of continental hydrologic droughts. *Hydrology papers (Colorado State University); no. 23*.

Yue, S. and Wang, C. Y. (2002). Applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test. *Water Resources Research*, 38(6):1068.

Zhang, J. (2004). Risk assessment of drought disaster in the maize-growing region of Songliao Plain, China. *Agriculture, ecosystems & environment*, 102(2):133–153.

A. COPYRIGHT AND CO-AUTHOR PERMISSIONS

Figure A.1. Copyright permission for Chapter 2

Requesting permission to reproduce WACE 2016 paper

Ganeshchandra Mallya <gmallya@purdue.edu> Fri, Mar 27, 2020 at 3:14 PM
To: Vimal mishra <vmishra@iitgn.ac.in>, Shivam <shiva@iitk.ac.in>, "Niyogi, Dev" <dniyogi@purdue.edu>, "Govindaraju, Rao S" <govind@purdue.edu>

Dear Co-authors:
I am including our paper titled *Trends and variability of droughts over the Indian monsoon region* in my thesis.
Can you kindly respond to this email permitting me to use this paper?
Thanks,
Ganesh

Citation:
Mallya, G., Mishra, V., Niyogi, D., Tripathi, S., and Govindaraju, R. S. (2016). Trends and variability of droughts over the Indian monsoon region. Weather and
Climate Extremes, 12:43-68.

Govindaraju, Rao S <govind@purdue.edu> Fri, Mar 27, 2020 at 3:15 PM
To: Ganeshchandra Mallya <gmallya@purdue.edu>, Vimal mishra <vmishra@iitgn.ac.in>, Shivam <shiva@iitk.ac.in>, "Niyogi, Dev" <dniyogi@purdue.edu>

Ganesh: Fine by me.

Best.

RSG

Niyogi, Dev <dniyogi@purdue.edu> Fri, Mar 27, 2020 at 3:32 PM
To: Ganeshchandra Mallya <gmallya@purdue.edu>, Vimal mishra <vmishra@iitgn.ac.in>, Shivam <shiva@iitk.ac.in>, "Govindaraju, Rao S" <govind@purdue.edu>

Yes please use it

Vimal mishra <vmishra@iitgn.ac.in> Fri, Mar 27, 2020 at 5:48 PM
To: "Niyogi, Dev" <dniyogi@purdue.edu>
Cc: Ganeshchandra Mallya <gmallya@purdue.edu>, "Govindaraju, Rao S" <govind@purdue.edu>, Shivam <shiva@iitk.ac.in>

Please go ahead and use them

Shivam <shiva@iitk.ac.in> Fri, Mar 27, 2020 at 9:57 PM
To: Ganeshchandra Mallya <gmallya@purdue.edu>
Cc: "Govindaraju, Rao S" <govind@purdue.edu>, "Niyogi, Dev" <dniyogi@purdue.edu>, Vimal mishra <vmishra@iitgn.ac.in>

Dear Ganesh:
Fine with me too.
Thanks,
Shivam

Figure A.2. Co-author permission for Chapter 2

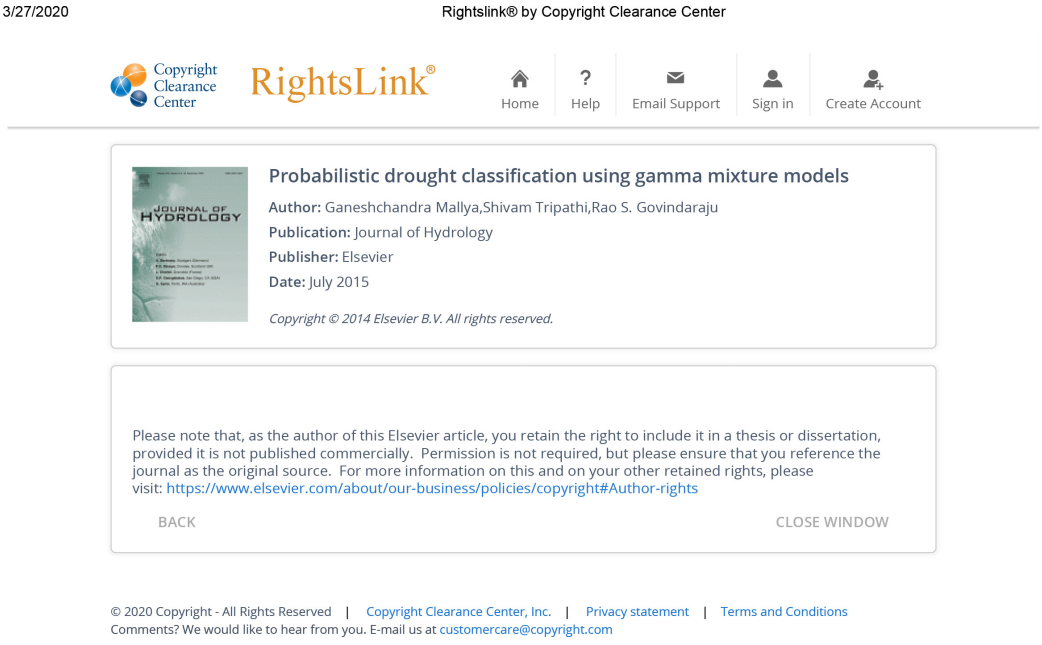

Figure A.3. Copyright permission for Chapter 3

Requesting permission to reproduce JoH (2015) paper
3 messages

Ganeshchandra Mallya <gmallya@purdue.edu> Fri, Mar 27, 2020 at 3:19 PM
To: Shivam <shiva@iitk.ac.in>, "Govindaraju, Rao S" <govind@purdue.edu>

Dear Co-authors:

I am including our paper titled *Probabilistic drought classification using gamma mixture models* in my thesis. Can you kindly respond to this email permitting me to use this paper?

Thanks,
Ganesh

Citation:
Mallya, G., Tripathi, S., and Govindaraju, R. S. (2015). Probabilistic drought classification using gamma mixture models. Journal of Hydrology, 526:116-126.

Govindaraju, Rao S <govind@purdue.edu> Fri, Mar 27, 2020 at 3:21 PM
To: Ganeshchandra Mallya <gmallya@purdue.edu>, Shivam <shiva@iitk.ac.in>

Ganesh:

That will be fine.

Best.

RSG

[Quoted text hidden]

Shivam <shiva@iitk.ac.in> Fri, Mar 27, 2020 at 9:59 PM
To: Ganeshchandra Mallya <gmallya@purdue.edu>, "Govindaraju, Rao S" <govind@purdue.edu>

Dear Ganesh:
Please use the paper.
Thanks,
Shivam
[Quoted text hidden]

Figure A.4. Co-author permission for Chapter 3

VITA

Ganeshchandra Mallya hails from Mangalore (Karnataka), India. He received his undergraduate degree in Civil Engineering from National Institute of Technology Karnataka (NITK), Surathkal, India. He then went on to work as a Systems Engineer in Tata Consultancy Services, Ltd. He received his M. S. in Civil Engineering from Purdue University, and then continued on for a Ph. D. degree.