January 2015

Probabilistic Models for Droughts: Applications in Trigger Identification, Predictor Selection and Index Development

Meenu Ramadas
Purdue University

Graduate School Form 30
Updated 1/15/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Meenu Ramadas

Entitled

PROBABILISTIC MODELS FOR DROUGHTS: APPLICATIONS IN TRIGGER IDENTIFICATION, PREDICTOR SELECTION AND INDEX DEVELOPMENT

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Dr. Rao S Govindaraju
Chair

Dr. Indrajeet Chaubey

Dr. Dev Niyogi

Dr. Venkatesh Merwade

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Dr. Rao S Govindaraju

Approved by: Dr. Dulcy Abraham 4/20/2015

Head of the Departmental Graduate Program Date

PROBABILISTIC MODELS FOR DROUGHTS: APPLICATIONS IN TRIGGER IDENTIFICATION, PREDICTOR SELECTION AND INDEX DEVELOPMENT

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Meenu Ramadas

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2015

Purdue University

West Lafayette, Indiana

For My Parents & Dear Teachers

ACKNOWLEDGEMENTS

I would like to express my special appreciation and sincere gratitude to my advisor Dr. Rao Govindaraju, for his endearing support and guidance throughout my Ph.D. at Purdue University. He has always been a strong source of motivation to me. I am also thankful to professors Dr. Indrajeet Chaubey, Dr. Dev Niyogi, and Dr. Venkatesh Merwade for serving as my dissertation committee members and for their insightful suggestions and valuable guidance during my Ph.D. I would also like to acknowledge the support and encouragement I have received from the faculty members of the Hydraulics and Hydrology group - Dr. Dennis Lyn and Dr. Cary Troy. I am grateful to Dr. Pradeep Mujumdar, Professor, IISc Bangalore, for his valuable guidance and encouragement extended to me. I am thankful to Dr. Rajib Maity, Associate Professor, IIT Kharagpur, who has been a very good mentor during my Ph.D. I am also grateful to the Lyles School of Civil Engineering office staff for their help and support.

I owe a big thanks to my parents - Mrs. Hemaletha N and Mr. Ramadas K, for their loving support and having let me pursue my academic goals in this far away land. It is beyond words how grateful I am to them for all their sacrifices. I would also like to express my gratitude to my brother and my grandparents. Their countless blessings and prayers have sustained me both personally and professionally, all these years.

Special thanks to my roommates Richa Ojha and Pikee Priya for the loveliest company and tremendous support during my stay in West Lafayette. I would also like to thank all of my G107 lab mates: Dr. Sultan Ahmed, Dr. Yamen Hoque, Dr. Jun Choi, Dr. Kuk-Hyun Ahn, Dr. John Newton, Becca Essig, Jessica Holberg, Sayan Dey, Adnan Rajib, Siddharth Saksena, Kyungmin Sung, Liuying Du, Zhu Liu, Keighobad Jafarzadegan, David Cannon and Barnard Mondal for their constant support. I am grateful to my special ones: Prashant Pendyala, Ganesh Mallya, Nikhil Sangwan, Shyamala Venkitaraman, Renu Dalal, Sampa Das, Saahas Bhardwaj, Jijo Mathew, Anup Mohan, Mayur Shindekar, Sonali Pattanayak, Dhanya C T, Ajithkumar K, Arun Sankar, Cibin Raj, Xiangning Huang and Sandeep Sasidharan, who have motivated me to work hard towards my goals, and have been with me during the ups and downs of my grad life. I am also thankful to all my friends in India for their constant motivation and love.

It has always been a privilege to be near and dear to many good friends and great mentors. I am glad that I have this opportunity to express my sincerest gratitude to the wonderful people I have in my life.

TABLE OF CONTENTS

Page

Page

Page

Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

LIST OF TABLES

Table Page

Table Page

ABSTRACT

Ramadas, Meenu. Ph.D., Purdue University, May 2015. Probabilistic Models for Droughts: Applications in Trigger Identification, Predictor Selection and Index Development. Major Professor: Rao S. Govindaraju.

The current practice of drought declaration (US Drought Monitor) provides a hard classification of droughts using various hydrologic variables. However, this method does not yield model uncertainty, and is very limited for forecasting upcoming droughts. The primary goal of this thesis is to develop and implement methods that incorporate uncertainty estimation into drought characterization, thereby enabling more informed and better decision making by water users and managers. Probabilistic models using hydrologic variables are developed, yielding new insights into drought characterization enabling fundamental applications in droughts.

Drought triggers are patterns in hydro-climatic variables that herald upcoming droughts and form the basis for mitigation plans. This thesis describes a new method for identification of triggers for hydrologic droughts by examining the association between the various hydro-climatic variables and streamflows over two study watersheds in Indiana, USA. The method combines the strengths of principal component analysis (PCA) for dimensionality reduction and copulas for building joint dependence. The expected values and ranges of predictor hydro-climatic variables for different streamflow quantiles are utilized to develop drought triggers for one-month lead time.

Accurate prediction of droughts requires a clear understanding of the dependence patterns among various influencing hydro-climatic variables and streamflows. A graphical modeling technique, employing conditional independence, is proposed to quantify the interrelationships between streamflows and a suite of available hydro-climatic variables, and to identify a reduced set of relevant variables for parsimonious model development. The graphical modeling approach is compared to the state-of-the-art method for predictor selection based on partial mutual information. For both a synthetic benchmark non-linear dataset and a watershed in southern Indiana, USA, this approach shows more discriminating results while being computationally efficient. The parsimonious models performed equally well as the models with the full set of original predictors.

In agricultural drought studies, soil moisture in the root zone of the soil is predominantly used to characterize agricultural droughts, but crop needs are rarely factored into the analysis. Accounting for crop responses to soil water deficits will provide a better representation of agricultural droughts, and is investigated in this thesis using crop stress functions available in the literature. A new probabilistic agricultural drought index is then developed within a graphical model (hidden Markov model) framework. This new index allows probabilistic classification of the drought states while taking into account the stress experienced by the crop due to soil moisture deficit. The method identified critical drought events and several drought occurrences that were not detected by popular indices such as standardized precipitation evapotranspiration index (SPEI) and self-calibrating Palmer drought severity index (SC-PDSI), and shows promise as a tool for agricultural drought studies.

An understanding of the role of hydrologic variables, either singly or in combination, is useful for assessment of overall drought status over a region. A multivariate cumulative density function (CDF)-based index is constructed using copulas, and probabilistic drought classification is performed using hidden Markov models. The resulting drought indices with various combinations of hydrologic variables are utilized to understand the roles of hydrologic variables for integrated drought assessment at watershed scales. In this thesis, the methodology is demonstrated using streamflow, precipitation and soil moisture variables to develop univariate and multivariate CDF-based indices at 1-, 3- and 6-month time scales. Drought characterization varied across the univariate, bivariate and trivariate drought models in the case study. Results are found to be watershed specific, and multivariate models tend to better capture the early onset of drought events and persistence of the drought states.

CHAPTER 1. INTRODUCTION

1.1 Background

Drought, as a prolonged status of water deficit, is perceived as one of the most expensive and the least understood natural disasters. In monetary terms alone, a typical drought costs American farmers and businesses $6-8 billion dollars each year, more than damages incurred from floods and hurricanes [FEMA, 1995]. The consequences tend to be more severe in areas where agriculture is a major economic driver. Dracup et al. [1980] stated that proper definition of drought depends on the nature of water deficit relevant to the study area. More than 150 definitions of droughts exist including both conceptual and operational definitions [Wilhite and Glantz, 1985].

Broadly, droughts have been classified into meteorological, agricultural, hydrologic and socio-economic droughts [Wilhite and Glantz, 1985; Mishra and Singh, 2010]. As water moves through the various components of the hydrologic cycle, precipitation deficits (meteorological droughts) lead to low soil moisture levels (agricultural droughts) that translate into low streamflows, reservoir and/or groundwater levels (hydrologic droughts). Drought conditions have a huge impact on allocation of resources, and hence affect the socio-economic status of dependent areas [Alcamo et al., 2007; Burn et al., 2008].

Drought assessment has long been conducted by comparing current conditions of different variables related to the aforementioned types of droughts to their long-term averages, with the magnitude of the deficit reflecting severity of the drought. Variables such as precipitation, soil moisture, streamflow, snowpack, water storage and availability, evaporation and crop production, are valuable entities in drought studies. A drought index, on the other hand, has the information derived by comparing current conditions to historical conditions or long term averages expressed using statistical formulae, providing a measure for quantifying droughts and their magnitude [Fuchs, 2014]. Palmer drought severity index [PDSI; Palmer, 1965], crop moisture index [CMI; Palmer, 1968], standardized precipitation index [SPI; McKee et al., 1993], soil moisture drought index [SMDI; Hollinger et al., 1993], vegetation condition index [VCI; Liu and Kogan, 1996], surface water supply index [SWSI; Shafer and Dezman, 1982], and reclamation drought index [RDI, developed as a part of the Reclamation States Drought Assistance Act of 1988] are some of the popular drought indices currently in use. They provide information on the major attributes of droughts namely the intensity, duration, severity and spatial extent. Each index has its advantages and limitations, and may be suitable for a specific application. Efforts to develop drought indices capable of addressing the probable causes/impacts of droughts have been underway for several decades [Panu and Sharma, 2002]. Existing practices of drought characterization (for instance, the United States Drought Monitor) follow a hard classification system using popular drought indicators. This methodology is, however, limited by a serious disadvantage of not being able to account for uncertainty in drought categorization.

A clear distinction can be made between a drought indicator, a trigger and an index. A drought trigger is the specific value of a drought indicator that dictates the onset and retreat of a drought, and determines the need for management and mitigation [Steinemann et al., 2005]. This information, regardless of the type of drought, is useful for making drought management decisions. Triggers can be expressed as range of values of drought indicators leading to a particular magnitude of drought, that help plan the timing of the response, and magnitude of damage expected. Long records of drought episodes that can be identified from historical records of drought-related variables and associated drought indicator values are required to develop drought triggers for any spatial location and at any time scale. Unlike drought indices that are defined, identification of drought triggers is recognized as a very challenging problem [Palmer et al., 2002].

Hydrologic variables are linked in complex ways. Precipitation and evaporation are acknowledged drivers of streamflows [Najjar, 1999; Chen et al., 2012]. In addition, soil moisture affects streamflow generation by controlling the partitioning of rainfall into runoff and infiltration [Western et al., 1999; Aubert et al., 2003]. Soil moisture possesses an intrinsic memory longer than several weeks to months [Entin et al., 2000; Koster et al., 2010], and hence, including soil moisture enhances hydrological modeling at seasonal lead times [Anctil et al., 2008; Mahanama et al., 2008]. Variables such as temperature, pressure and wind speed are also important, as they control evapotranspiration losses and subsequently the amount of soil moisture. Surface air temperature, evaporation and mean sea level pressure are known to influence the magnitude and occurrence of rainfall over a region, and consequently streamflows [Ward, 1992; Parthasarathy et al., 1993; Trenberth,

1999]. Researchers rely on models to improve drought predictions using these variables. However, including all the predictors in the model increases the dimensionality of the problem, and does not always guarantee the best prediction results. The knowledge of interdependencies between variables could be utilized to include only the relevant predictors to yield parsimonious hydrological models. Predictor selection is therefore an integral component of the development of prediction models for streamflows and hydrological droughts. Among these, data-driven algorithms have been found to possess computational ease and robustness in predictor identification and model development.

The droughts of the 1930s, 1950s, 1980s and 1990s in the last century in the United States had significant impact on the agricultural sector [Narasimhan and Srinivasan, 2005]. The most recent 2012 Midwest drought in the US severely affected the agricultural activities across the Corn Belt [Elliot et al., 2013] and the Midwest states [Mallya et al., 2013a]. Al Kaisi et al. [2013] conducted a detailed study of the unfavorable soil conditions and changes in soil strata in the state of Iowa as a result of the 2012 drought. The authors state that changing soil water relationships could have detrimental effects on cultivation. Agricultural droughts develop when soil moisture deficits adversely affect crop growth, health, and yields, and are aggravated by periods of inadequate irrigation. They are characterized by lack of soil moisture, driven by prolonged periods of precipitation deficits, and followed by adverse effects on crop productivity [Heim, 2002; Wilhite, 2005]. Meteorologic and hydrologic drought indices (e.g., SPI and PDSI) have been often used in agricultural drought studies [Narasimhan and Sreenivasan, 2005]. The PDSI uses both precipitation and surface air temperature as inputs, in contrast to SPI that uses precipitation alone. However, PDSI is limited as an

indicator of soil moisture status or as being capable of identifying agricultural droughts; it demonstrates good correlation with soil moisture content during warm seasons but weak correlation in spring as the underlying model does not account for the effect of snowmelt [Dai et al., 2004]. Palmer [1968] developed the crop moisture index (CMI) as an index for short-term agricultural droughts from procedures similar to the PDSI. The CMI is computed from evapotranspiration deficits for monitoring short-term agricultural drought conditions that modulate crop growth. Meyer et al. [1993] developed a crop specific drought index (CSDI) for corn using evapotranspiration estimates. An alternative drought index standardized precipitation evapotranspiration index (SPEI)−that possesses the merits of PDSI and SPI in terms of sensitivity to temperature-driven evaporation that is important in crop growth and multi-scalar properties, respectively, was proposed by Vicente-Serrano et al. [2010]. The performance of SPEI in drought impact analyses and climate change studies is well documented [Yu et al., 2013; Potop et al., 2012; Vicente-Serrano et al., 2010]. However, studies in the past have not addressed crop water stress-based drought characterization schemes for agricultural droughts. There is a growing need for more research to understand and develop models/tools to monitor agricultural droughts. It is also desirable to design these models to account for uncertainty in drought classification.

An overall drought assessment model over a watershed requires that variables representing different types of droughts, namely hydrological, meteorological and agricultural droughts, be included in the analysis. Numerous studies have recommended multivariate drought indices with different choice of variables [Keyantash and Dracup, 2004; Karamouz et al., 2009; Vicente-Serrano et al.; 2010; Kao and Govindaraju, 2010;

Rajsekhar et al., 2014; Hao and Aghakouchak, 2014]. Drought characterization varies with different combinations of hydrologic variables present in the model. Among the vast suite of variables that drive droughts, a smaller subset if identified, could be used for efficiently performing overall drought monitoring and assessment. Previous studies have not directly addressed these aspects.

1.2 Motivation

Although probabilistic models exist for hydrological modeling and drought prediction and several indices have been designed for addressing drought assessment over the past century, these formulations are not suitable for development of triggers that require identification of ranges of predictor variables that herald a particular drought. Proven methodologies for parsimonious and robust models for drought analyses are lacking. Two major limitations are encountered in drought applications, namely the large dimensionality of predictor hydro-climatic variables, and modeling the joint dependence of predictands and relevant predictors. The motivation for this research is to develop and demonstrate the utility of probabilistic approaches to overcome these limitations, and bring uncertainty estimation into drought characterization thereby enabling informed decision making by water users and managers. This is accomplished by filling in some of the research gaps identified in the extraction of hydrologic drought triggers, predictor selection techniques for drought forecasting, developing probabilistic models for agricultural droughts, and the role of choice of hydrologic variables in multivariate drought monitoring.

1.3 Objectives of the Thesis

The objectives of this research are as follows:

i. To explore patterns in hydro-climatic variables as potential precursors to hydrologic droughts in watersheds.

The joint distribution of streamflows and the important principal components of precursor hydro-climatic variables is modeled using an appropriate copula family for two study watersheds in Indiana, USA. The PCA-copula framework is then utilized to develop drought trigger information. While copulas and PCA have been widely used individually, no prior studies exist for identifying drought triggers in this fashion.

ii. To extract the conditional independence structure between streamflow and prominent hydro-climatic variables, so as to develop a parsimonious multivariate statistical approach to streamflow/drought forecasting while honoring the dependence structure among the competing predictor variables.

A graphical model-based approach allows for predictor selection as well as development of a streamflow forecasting model. The efficacy of this approach for supervised predictor selection from a pool of interdependent variables has not been evaluated in hydrologic applications.

iii. To develop a probabilistic drought assessment model for agricultural droughts based on the concept of crop water stress using graphical models.

Using a crop water stress function rather than soil moisture data will allow for characterization of agricultural droughts based on crop needs. By taking into account the crop-specific soil moisture requirements, the drought categorization from this index will

be more reflective of crop needs. Graphical models, specifically hidden Markov models, are utilized for probabilistic classification using the proposed index.

iv. To explore the choice of hydrologic variables in overall drought monitoring at a watershed scale, over multiple time scales.

Different hydrologic variables could be combined to yield models for overall drought assessment. Drought evolution in the different models is studied to understand the roles of selection of variables for drought classification. Use of cumulative probabilities from joint cumulative density functions (CDFs) as drought indicators in a hidden Markov model (HMM) framework allow for probabilistic drought categorization.

1.4 Organization of the Thesis

The current chapter provides background and motivation for this study. In Chapter 2, the first objective, the identification and development of hydrological drought triggers is discussed in detail. The predictor selection problem for streamflows and hydrological droughts, i.e., the second objective, is described in Chapter 3. The methodology and results for the third objective, to develop a new agricultural drought index that accounts for crop water stress, are presented in Chapter 4. In Chapter 5, the results of a case study on multivariate probabilistic drought analysis at multiple time scales are discussed. Chapter 6 contains the summary and conclusions derived from the drought studies.

CHAPTER 2. IDENTIFICATION OF HYDROLOGIC DROUGHT TRIGGERS FROM HYDRO-CLIMATIC PREDICTOR VARIABLES

2.1 Abstract

Drought triggers are patterns in hydroclimatic variables that herald upcoming droughts and form the basis of mitigation plans. This chapter develops a new method for identification of triggers for hydrologic droughts by examining the association between various hydroclimatic variables and streamflows. Since numerous variables influence streamflows to varying degrees, principal component analysis (PCA) is utilized for dimensionality reduction in predictor hydroclimatic variables. The joint dependence between the first two principal components, that explain over 98% of the variability in the predictor set, and streamflows is computed by a scale-free measure of association using asymmetric Archimedean copulas over two study watersheds in Indiana, USA, with unregulated streamflows. The M6 copula model is found to be suitable for the data and is utilized to find expected values and ranges of predictor hydroclimatic variables for different streamflow quantiles. This information is utilized to develop drought triggers for 1 month lead time over the study areas. For the two study watersheds, soil moisture, precipitation, and runoff are found to provide the fidelity to resolve amongst different drought classes. Combining the strengths of PCA for dimensionality reduction and copulas for building joint dependence allows the development of drought triggers.

2.2 Introduction

The occurrence and magnitude of hydrologic droughts are heralded by triggers that may be manifested in specific patterns of hydro-climatic variables. Identification of these triggers at appropriate lead times is necessary for devising effective drought mitigation plans. Estimating water deficits and drought categories at weekly, monthly, seasonal, and annual lead times are needed for scheduling irrigation events and managing water resources of a region. Drought characterization is currently accomplished by indices such as Standardized Precipitation Index (SPI), Palmer Drought Severity Index (PDSI), Crop Moisture Index (CMI), Surface Water Supply Index (SWSI), and Reclamation Drought Index (RDI; developed as a part of the Reclamation States Drought Assistance Act of 1988). Drought indices are typically designed for assessing current conditions, and have little predictive capability. Large scale oceanic and atmospheric indicators such as the El Niño-Southern Oscillation (ENSO) phases, North Atlantic Oscillations (NAO), Pacific North American index (PNA), Atlantic Multidecadal Oscillations (AMO), and Pacific Decadal Oscillations (PDO) are used as long term precursors to annual/seasonal forecasts of precipitation [Ropelewski and Halpert, 1996; McHugh and Rogers, 2001; Maity and Kumar, 2008a]. However, for many parts of the world, including Indiana, USA, these indicators have been found to have little to no influence [Charusombat and Niyogi, 2011]. Further, their incapability to provide short-term predictions (several weeks, to 6-month range) render them unsuitable as drought triggers for such time scales. We hypothesize that hydrological droughts, reflected in unregulated streamflows, would have precursors in local hydro-meteorologic variables

related to rainfall and soil moisture over the corresponding watersheds. McKay et al. [1989] suggested that accurate drought predictions will need models that link between climate and weather factors to streamflows and river stage data.

Several considerations come into play for the development of drought triggers including drought types, data availability, choice of hydrologic variables (precipitation, temperature, streamflows, storage levels, etc.), temporal scales and validity of the trigger. Over the past two decades, drought triggers have been developed by several states and utilities [Steinemann, 2003]. However, these have met with limited success because of (i) anomalies between results from different drought indicators, and (ii) lack of a strong record length for proper model development and validation exercises. Moreover, these triggers are often defined as some preset thresholds to be crossed by various drought indices at the same instance of time for which drought status is being analyzed. Thus, they may not recognize early warning signals that may be present in the record.

Though droughts are fundamentally triggered by insufficient precipitation, the evolution of water deficits from precipitation to soil moisture and to streamflows is not instantaneous and is controlled by complex physical mechanisms. As hydrologic droughts are based on abnormally low flows, estimation of streamflows is therefore a necessary prerequisite to drought analysis. Since a drought trigger governs the level of future response, it is important that the trigger be based on methods that convey predictive uncertainty. There are many methods available for estimation of streamflows, classified mainly into physics-based, conceptual, and data-driven approaches. Several watershed models have been developed that rely upon the physical knowledge of the watershed and the hydrological cycle, often resulting in complex representations that

require intensive computer effort for model calibration and corroboration. Data-driven techniques do not require detailed understanding of the inherent physical mechanisms, but have shown comparable accuracy for streamflow prediction as physics-based models [Wu et al., 2009]. The time-scale of one-month lead forecasts is particularly challenging because physics-based models (HEC-HMS, MIKE-SHE, etc.) are not able to project using input data beyond several hours to days without a disaggregation procedure. Process-based models such as SWAT perform simulations at a daily time step [Srinivasan and Arnold, 1994], and model outputs have to be aggregated to obtain monthly values. However, the strength of such models lies in examining long-term consequences of management practices rather than monthly forecasts. There are many conceptual lumped-parameter models developed in the last four decades, mainly for flood forecasting, with one day or shorter time resolutions [Xu and Singh, 2004], but their predictive capabilities are very limited if the time horizon exceeds several days.

Statistical approaches have been utilized to model the complex relationships between streamflows and the large-scale atmospheric circulation phenomena [Anmala et al., 2000; Maity and Kumar, 2008b]. The predictors used in majority of these data-driven approaches were hydro-climatic variables such as mean temperature, mean sea level pressure, soil moisture, precipitation, runoff and wind speed. While these studies have stressed the importance of hydro-climatic variables for enhancing streamflow prediction, they were primarily targeted towards long-range forecasting [Salas et al., 2011]. Even with the predictor set identified, new approaches are needed for achieving short-term (few weeks to months) forecasts. The use of advanced statistical models based on Markov properties [e.g. Mallya et al., 2013] have helped in probabilistic classification of

drought states and alleviated the need for user-specified thresholds for drought categorization. Thus, though robust models exist for forecasting streamflows and upcoming hydrologic droughts, these models are not suitable for development of triggers that require identification of the ranges of predictor variables that herald a particular drought.

The joint probability density function between streamflows and hydro-climatic predictor variables is needed to identify and develop drought triggers. Copulas are a natural choice for this task [Nelsen, 2006]. They allow the dependence structure to be modeled without any restriction on the distributions of the marginals [Genest and Favre, 2007], and have been gaining popularity with hydrologic applications. Favre et al. [2004] used Frank and Clayton 2-copulas to model the dependence between streamflow peaks and volumes. Salvadori and De Michele [2004] adopted copulas in their study of the return period of hydrological events. Zhang and Singh [2006] used copulas to determine bivariate distributions between flood peaks, volumes and durations, and employed them to define joint and conditional return periods needed for hydrologic design calculations. The joint distribution of intensity, duration and severity of droughts was modeled using copulas by Shiau et al. [2007], Wong et al. [2010], and Madadgar and Moradkhani [2013]. Maity and Kumar [2008a] analyzed the dependencies among the teleconnected hydro-climatic variables using copulas for the prediction of response variables using large scale oceanic and atmospheric indicators. Kao and Govindaraju [2010a] utilized copulas to construct an inter-variable drought index, where the dependence structure of precipitation and streamflow marginals was preserved. The review by Mishra and Singh [2010] highlights the expanding role of copulas in drought assessment studies.

Given the large number of potential hydro-climatic variables in the predictor set, the direct use of copulas to model their joint dependence with streamflows is impractical because of the mathematical complexity in constructing higher-dimensional copulas. If the dependence between all the interacting variables cannot be represented by multivariate Gaussian (or meta-elliptical) copulas, then models at even the trivariate level can be very challenging [Kao and Govindaraju, 2008, 2010b]. Moreover, with multiple interacting variables, the curse of dimensionality adds further challenges to estimation of model parameters from limited record lengths. While many options exist for modeling bivariate dependence between variables, models for higher dimensions are not easily available.

Principal Component Analysis (PCA) provides an elegant way of projecting the precursor hydro-climatic variables onto a feature space, and representing the original data through a reduced number of effective features called principal components [Jolliffe, 1986; Preisendorfer, 1988]. If the first few (two in this case) features are able to explain most of the variability (>90%) in the original data set, then substantial dimensionality reduction may be achieved through unsupervised learning. PCA is recognized as the most widely used tool for dimensionality reduction for multivariate data problems. Lins [1985] utilized PCA to construct parsimonious models for multi-site streamflows. Maurer et al. [2004] showed the effectiveness of PCA for both reducing the dimensionality of large data sets and better graphical representation of the modes of variability in streamflows. Tripathi and Govindaraju [2008] developed algorithms for data compression using PCA for data sets with noise. PCA was adopted by Keyantash and Dracup [2004] to achieve dimensionality reduction for developing an aggregate drought index.

The goal of this chapter is two-fold. The first goal is to model the joint distribution of streamflows and the important principal components of precursor hydro-climate variables using an appropriate copula family for two study watersheds in Indiana, USA. This copula model is tested for its capability to forecast low streamflows that are of concern for hydrologic droughts. The second goal is to utilize the PCA-copula framework to develop drought trigger information. While copulas and PCA have been widely used individually, to the best of my knowledge, no prior studies exist for identifying drought triggers in this fashion. The details of study watersheds are provided in section 2.3. The methodology adopted in the study with details of principal components analysis (PCA), copula models and drought trigger analysis are explained in section 2.4. These are followed by results and discussion in section 2.5, and the summary and conclusions of the study in section 2.6.

2.3 Study Area and Data Used

2.3.1 Study Area

The study was carried out over two watersheds in the state of Indiana, USA. Both the watersheds form a part of the Ohio River Basin. The first watershed (WS I) extending from 38°34’N to 39°49’N and 85°24’W to 86°31’W spreads over 6259 square kilometers. The second watershed (WS II) lies between 40°47’N to 41°24’N and 85°8’W to 86°20’W and extends over an area of 1657 square kilometers. The two watersheds are shown in Figure 2.1. The land use in these watersheds consists of mainly agricultural and forest lands, followed by public and urban built-up lands. Agriculture being the major

economic activity prevalent in WS I and WS II, high irrigation water demands exist during the growing season. The choice of the watersheds was governed by the need to conduct drought analyses for locations, where streamflows were not influenced by human activities.

2.3.2 Data Used

The 30 m resolution DEMs obtained from USGS National Elevation Data set was used to delineate the watersheds. Though the choice of coarser resolution affects the identification of drainage features in low relief landscapes, there is substantial reduction in computational efforts involved in the processing of the 30 m digital elevation model (DEM) over a high-resolution DEM. Modeling the dependencies and analysis of drought triggers require a long record of historic observations. Therefore, monthly data with a minimum record length of 50 years were adopted in the present study. The various hydroclimatic variables used in the study are listed in Table 2.1. The 0.5° grid resolution climate prediction center (CPC) global monthly data sets [Huang et al., 1996; Fan and van den Dool, 2004], available from 1948 onwards, were used. The land model was treated as a one-layer 'bucket' water balance model, when generating the CPC data sets. The data used in our study include modeled monthly soil moisture values, modeled monthly runoff values, observed monthly precipitation values, observed monthly temperature values, and modeled monthly evaporation values. The location of CPC stations is marked by circles in Figure 2.1. Given the small watershed sizes determined by the need for unregulated streamflows, the number of CPC grid points directly over the study areas is quite small. The variables: sea-level pressure, u-wind, and v-wind were

obtained from the NCEP/NCAR Reanalysis-1 project data, at a spatial resolution of 2.5° X 2.5° [Kalnay et al., 1996]. The resultant of the u-wind and v-wind components was adopted as the wind speed variable in the present study. Given the monthly time scale chosen for this study, the time of concentration for these watersheds is in the order of days. Thus, variables were multiplied by the Thiessen weights at different grid points to obtain their spatially averaged values over the study watersheds. The US Geological Survey (USGS) monthly streamflow data from 1958 to 2010 recorded at the USGS 03371500 (East Fork White River near Bedford, Indiana) were used for WS I, while the data at USGS streamflow gage 03328500 (Eel River near Logansport, Indiana) from 1948 to 2010 were used for WS II.

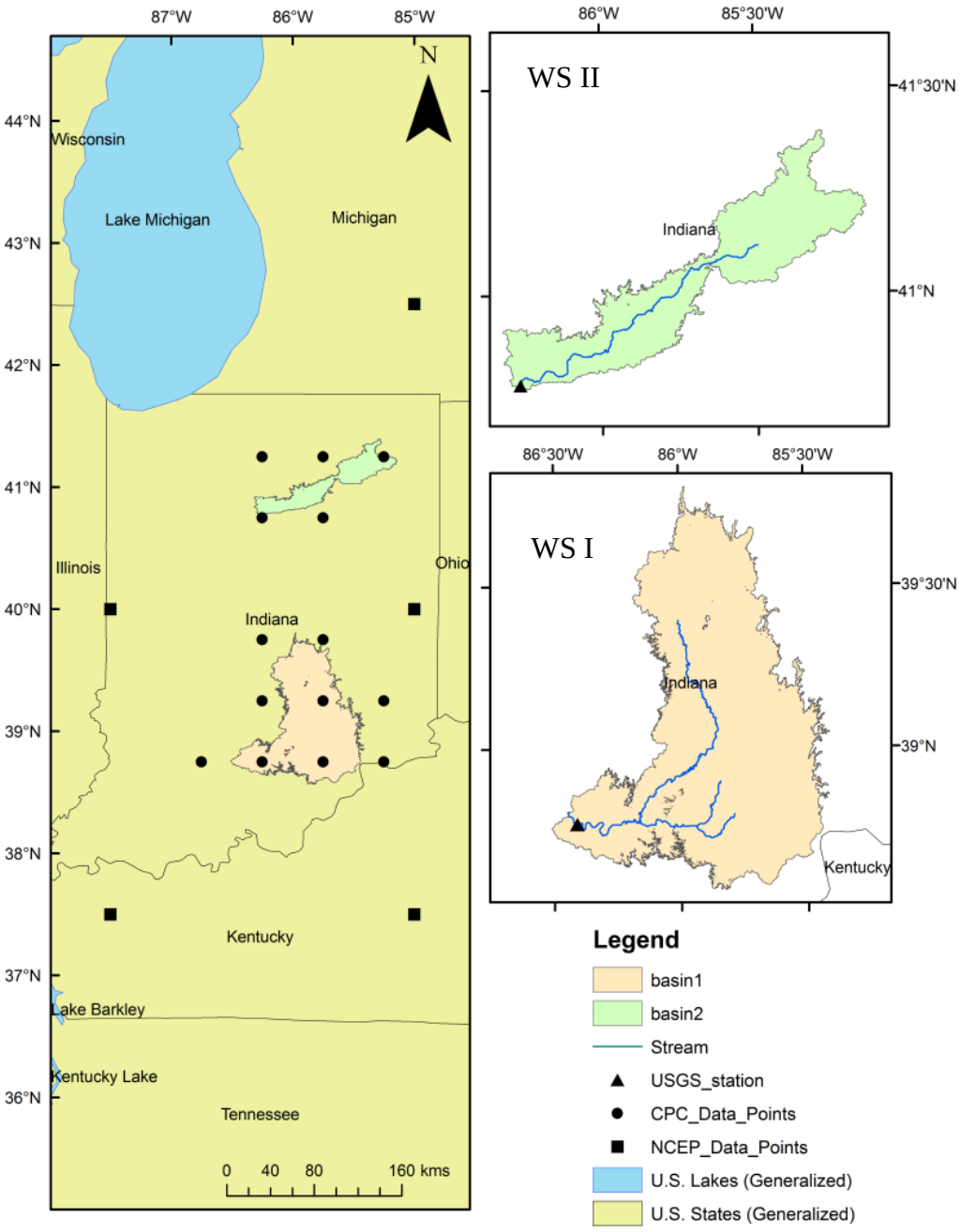

Figure 2.1 Map of the study watersheds WS I and WS II

Table 2.1 List of variables used in the study

Sl. No	Variables Used	Abbreviation	Unit
1	Soil moisture	SMTR	mm
2	Precipitation	PPTN	mm
3	Temperature	TEMP	°C
4	Runoff	RNF	mm
5	Evaporation	EVPN	mm
6	Sea level pressure	PSSR	mbar
7	Wind Speed	WIND	m/s
8	Streamflow	SF	m^3/s

2.4 Methodology

2.4.1 Dimensionality Reduction Using Principal Components Analysis

The formulation of a dependence model between the seven predictor variables in Table 2.1 and streamflows is impractical even when using copulas. PCA was performed to transform the set of correlated *n*-dimensional (*n*=7 here) predictor set into another set of *n*-dimensional uncorrelated vectors (called principal components). The PCs are arranged in order of their ability to explain the variability in the data. The conventional or standard PCA, which is formulated as an eigenvalue problem, was used for unsupervised dimensionality reduction [Jolliffe, 1986]. Prior to extracting the principal components, the mean value was subtracted from each of the predictors to obtain a series of predictor anomalies. The covariance matrix was obtained for the anomaly data sets, and the eigenvalues and eigenvectors of this covariance matrix were computed. The degree of dimensionality reduction achieved in the predictor set was determined by variance explained by the first two principal components.

2.4.2 Asymmetric Archimedean Class of Copulas

A copula is a function that models the dependence between multiple random variables, regardless of their marginals. A d-dimensional copula is a multivariate cumulative density function (CDF) C defined in the unit d-dimensional space $[0,1]^d$ with uniform margins $[0,1]$ and with the following properties: (i) $\forall u \in [0,1]^d, C(u) = 0$ if at least one coordinate of u is equal to 0, and $C(u) = u_k$ if all the coordinates of u are equal to 1 except u_k; (ii) $\forall a$ and $b \in [0,1]^d$ such that $a \le b$, $V_c([a,b]) \ge 0$, where V is the C-volume [Nelsen, 2006]. The copula approach to dependence modeling has its roots in the theorem by Sklar [1959], according to which a d-dimensional CDF with univariate margins $F_1, F_2, ..., F_d$ is defined by

$$H(x_1, x_2, .., x_d) = C(F_1(x_1), F_2(x_2), ..., F_d(x_d)) = C(u_1, u_2, ..., u_d) \tag{2.1}$$

where $F_k(x_k) = u_k$ for $k = 1, 2, ..., d$ with $U_k \in U(0,1)$ if F_k is continuous.

Archimedean copulas are very popular, with both symmetric and asymmetric forms available in the literature [Joe, 1997; Nelsen, 2006]. They possess closed form expressions and allow modeling of a variety of different dependence structures. An Archimedean symmetric d-copula is of the form

$$C(u) = \varphi^{-1}\left(\sum_{k=1}^{d} \varphi_k(u_k)\right) \tag{2.2}$$

where the function φ (called the generator of the copula) is a continuous strictly decreasing function from $[0,1]$ to $[0,\infty)$, such that $\varphi(0) = \infty$ and $\varphi(1) = 0$, and its

inverse φ^{-1} is completely monotone on $[0,\infty)$ i.e., φ^{-1} has derivatives of all orders which alternate in sign [Nelsen, 2006]:

$$(-1)^k \cdot \frac{d^k \varphi^{-1}(t)}{dt^k} \geq 0 \tag{2.3}$$

for all t in $[0,\infty)$ and $k = 1,2,...,d$.

In equation (2.2), if a certain u_k is assigned the value 1, then the joint distribution of $(u_1,u_2,...,u_d \mid u_k)$ is obtained. Since $\varphi(u_k) = 0$ when $k = 1$, the $(d-1)$-dimensional marginal of the symmetric Archimedean copula is also an Archimedean copula. The expressions for these $(d-1)$-dimensional copulas are identical regardless of the choice of k. As a result, only one Archimedean 2-copula is required to model all mutual dependencies among the variables. This exchangeability property that can be modeled by symmetric copulas limits the nature of the dependence structures. Since the study took into account correlated variables such as streamflows and principal components that possess different bivariate dependence structures, a more general multivariate extension of the Archimedean 2-copula, namely the fully nested or asymmetric copula as described in Whelan [2004], was adopted here. This copula is given by $d-1$ distinct generating functions as:

$$C(u_1,u_2,...,u_d) = C_1\left(u_d, C_2\left(u_{d-1},...,C_{d-1}(u_2,u_1)...\right)\right) \tag{2.4}$$

For example, in a fully nested 3-copula, two variables u_1 and u_2 are coupled using copula C_2 and the copula of u_1 and u_2, is coupled with u_3 by copula C_1. In general, there are $d(d-1)/2$ ways of coupling d variables. When the bivariate joint probability

of two variables conditioned on the third variable is computed, different dependence structures are obtained based on the conditioning variable. Grimaldi and Serinaldi [2006] used asymmetric Archimedean copulas to model trivariate joint distribution of flood peaks, volumes and durations. A nested 3-copula was adopted in the present study to model the dependence between the monthly streamflow anomaly and the first two principal components of a set of predictor variables. There are two parameters for the nested 3-copula model: θ_1 and θ_2 such that $\theta_1 \leq \theta_2$ implying a higher degree of dependence for the inner nested variables. It has been found that only two dependence structures can be reproduced for three possible pairs [Grimaldi and Serinaldi, 2006]. When two variables u_1 and u_2 are likely correlated with the third one u_3, and the degree of dependence between u_1, u_2 is stronger than that of either u_1 and u_2 with u_3, the asymmetric 3-dimensional model may be applied. The dependence between the variables is expressed in terms of the Kendall's correlation coefficient, τ. Kendall's τ for a random vector $(X,Y)^T$ is simply the probability of concordance minus the probability of discordance [Embrechts et al., 2003]:

$$\tau_{XY} = \text{Prob}\left\{\left(X-\tilde{X}\right)\left(Y-\tilde{Y}\right)>0\right\} - \text{Prob}\left\{\left(X-\tilde{X}\right)\left(Y-\tilde{Y}\right)<0\right\} \quad (2.5)$$

The various asymmetric Archimedean copula families selected for the study, their permissible θ values, and dependence ranges are listed in Table 2.2.

2.4.3 Parameter Estimation

Several copula parameter estimation methods are available in the literature namely, the method of moments, canonical maximum likelihood method, and inference from margins method. When one-parameter bivariate copulas are adopted, the popular approach is the simple method of moments based on inversion of Spearman's or Kendall's rank correlation [Genest and Favre, 2007]. In the multivariate-multiparameter case, this method becomes less elegant and may lead to inconsistencies. In such instances, a more natural estimation technique is the canonical maximum likelihood (CML) method [Genest et al., 1995; Kojadinovic and Yan, 2011]. The parameters of the five nested 3-copula families used in this study were estimated using the CML method. This method performs a non-parametric estimation of the marginals by using the respective scaled ranks. The dependence parameters θ_1 and θ_2 are obtained by maximizing the log-likelihood function $l(\theta)$ given by:

$$l(\theta) = \sum_{i=1}^{n} \log\left[c_\theta \left\{ \hat{F}_1(x_{i1}), \hat{F}_2(x_{i2}), \ldots, \hat{F}_d(x_{id}) \right\} \right] \tag{2.6}$$

where c_θ denotes the density of the copula C_θ, and $\hat{F}_k(x_{ik})$ (also denoted as u_k) is the rank-based non-parametric marginal probability of k^{th} variable given by:

$$\hat{F}_k(x_{ik}) = \frac{1}{n+1} \sum_{i=1}^{n} I(X_{ik} \le x_{ik}) \qquad k = 1, 2, ..., d \tag{2.7}$$

where $I(\bullet)$ is indicator function returning 1 if the argument is true and 0 otherwise.

Table 2.2 Asymmetric Archimedean copula families used in the study

Type	Nested Copula $C_{\theta_1}\left(u_3, C_{\theta_2}\left(u_1, u_2\right)\right)$	$\theta_2 \geq \theta_1 \in$	$\tau_{12}, \tau_{23}, \tau_{13} \in$	Reference
M3	$-\theta_1^{-1} \log \{1-(1-e^{-\theta_1})^{-1}(1-[1-\left(1-e^{-\theta_2}\right)^{-1}(1-e^{-\theta_2 u_1}))(1-e^{-\theta_2 u_2})]^{(\theta_1/\theta_2)})(1-e^{-\theta_1 u_3})\}$	(0,∞)	(0,1)	Joe, 1997
M4	$[(u_1^{-\theta_2}+u_2^{-\theta_2}-1)^{(\theta_1/\theta_2)}+u_3^{-\theta_1}-1]^{(-1/\theta_1)}$	(0,∞)	(0,1)	Joe, 1997
M5	$1-\left[\left\{\left(1-u_1\right)^{\theta_2}\left(1-\left(1-u_2\right)^{\theta_2}\right)+\left(1-u_2\right)^{\theta_2}\right\}^{(\theta_1/\theta_2)}\left(1-\left(1-u_3\right)^{\theta_1}\right)+\left(1-u_3\right)^{\theta_1}\right]^{(1/\theta_1)}$	(1,∞)	(0,1)	Joe, 1997
M6	$e^{\{-([(-\log u_1)^{\theta 2}+(-\log u_2)^{\theta 2}]^{(\theta 1/\theta 2)}+(-\log u_3)^{\theta 1})^{(1/\theta 1)}\}}$	(1,∞)	(0,1)	Joe, 1997; Embrechts et al., 2003
M12	$\{([(u_1^{-1}-1)^{\theta_2}+(u_2^{-1}-1)^{\theta_2}]^{(\theta_1/\theta_2)}+(u_3^{-1}-1)^{\theta_1})^{(1/\theta_1)}+1\}^{-1}$	(1,∞)	(0.333,1)	Embrechts et al., 2003

2.4.4 Goodness-of-fit Tests for Asymmetric Copulas

When there exist more than one feasible copula families that satisfy the dependence range for the given data, the final selection of a suitable copula is based on the best fit to observations. This fit can be assessed graphically by comparing the scatter plots of observed and simulated data in the case of bivariate distributions, but becomes difficult for higher dimensions. Goodness-of-fit tests examine the null hypothesis $H_0 : C \in C_0$ for a copula class C_0 against $H_1 : C \notin C_0$. These tests compare the distance between the empirical distribution of copula, C_n and an estimation of C_{θ_n} of C obtained under H_0 [Genest et al., 2009]. Formally, the goodness-of-fit tests are based on the statistic:

$$\Omega = \sqrt{n}\left\{C_n(u) - C_{\theta_n}(u)\right\} \quad u \in [0,1]^d \tag{2.8}$$

where the empirical copula of the data $X_1, X_2, ..., X_d$ is defined by Deheuvels [1981] as:

$$C_n(u) = \frac{1}{n}\sum_{i=1}^{n} I\left(U_i \le u\right), u \in [0,1]^d \tag{2.9}$$

In this study, the rank-based versions of Cramér-von Mises and Kolmogorov-Smirnov statistics were used for testing the goodness-of-fit of the nested copulas. The Cramér-von Mises statistic S_n has been a popular goodness-of-fit test procedure for copula models [Genest et al., 2009]. The statistic S_n was determined using Equation (2.10), using C_n, the empirical copula computed as per Equation (2.9), and substituting the value of C_θ evaluated from the copula expression:

$$S_n = \sum_{i=1}^{n} \left\{ C_n\left(\hat{U}_i\right) - C_{\theta_n}\left(\hat{U}_i\right) \right\}^2 \tag{2.10}$$

The Kolmogorov-Smirnov statistic T_n utilizes the absolute maximum distance between the empirical copula probability distribution and that simulated using the estimated parameters to measure the fit of the copulas as shown below [Genest et al., 2009].

$$T_n = \max_{u \in [0,1]^d} \left| \sqrt{n} \left\{ C_n(u) - C_{\theta_n}(u) \right\} \right| \tag{2.11}$$

Additionally, the probability plots of the empirical distribution and the nested copula families were compared to assess the performance of copulas. The family providing the best fit based on the above criteria was selected for subsequent analysis.

2.4.5 Streamflow Forecasting and Drought Analysis

The joint dependence modeled using the best copula was employed to estimate 1 month ahead streamflows. The probabilistic predictions of streamflows at different quantiles were made using the copula function. The expected values of monthly streamflows during the model development and model testing periods were computed. The range of forecasts was quantified by estimating predictions at 2.5% and 97.5% probabilities, i.e., 95% confidence interval for the prediction. The forecasts of streamflow were analyzed to identify the occurence of extremes, particularly for droughts in the study area. Given the focus on streamflows in this study, hydrological droughts were characterized by the standardized streamflow index that is similar to the SPI introduced by McKee et al. [1993] for meteorological drought analysis. The long-term streamflows

record was fitted to a gamma probability distribution and then transformed to a standard normal distribution through the quantiles so that the mean standardized index for a certain location and particular period (1 month) is zero [Edwards and McKee, 1997]. A positive value of the index shows the degree of wetness, while a negative value indicates the severity of streamflow deficit. The ranges of this drought index for different hydrological conditions, labeled exceptionally dry (D4) to exceptionally wet (W4), are presented in Table 2.3. This drought severity classification based on SPI values was adopted from http://droughtmonitor.unl.edu/classify.htm. The streamflows estimated using copula were used for the prediction of droughts in the study areas.

Table 2.3 Range of drought index for different hydrological states

State	Description	Drought Index
D4	Exceptional drought	-2 or less
D3	Extreme drought	-1.6 to -1.9
D2	Severe drought	-1.3 to -1.5
D1	Moderate drought	-0.8 to -1.2
D0	Abnormally dry	-0.5 to -0.7
Normal	Normal condition	-0.4 to 0.4
W0	Abnormally wet	0.5 to 0.7
W1	Moderately wet	0.8 to 1.2
W2	Severely wet	1.3 to 1.5
W3	Extremely wet	1.6 to 1.9
W4	Exceptionally wet	2 or more

2.4.6 Analysis for Drought Triggers

The occurrence of hydrological extremes in the study areas was highly correlated with the local hydroclimatic variables at 1 month lead times, and as such short-term predictions of droughts could be achieved. The joint dependence information contained in

the copula was exploited to obtain the expected values of the climate precursor anomalies conditioned on a streamflow anomaly. This allowed for identification of patterns in the precursors that could trigger hydrological droughts of different categories.

2.5 Results and Discussion

2.5.1 Principal Components Analysis

The anomalies of hydroclimatic predictors and streamflows at monthly scale were obtained by subtracting their respective monthly means. The dependence between the first two principal components of the anomalies of these variables was represented by a joint asymmetric copula in the present study and was used to predict streamflows. The data from January 1958 to December 1993 were used for developing the statistical model for WS I, whereas model development period for WS II was from January 1948 to December 1990. Thus, two thirds of the data were used for model training and the remainder used for evaluating model performance.

Starting from the large suite of potential predictors, PCA was used for dimensionality reduction. The results of principal components analysis performed on the predictor variables for the two watersheds are given in Table 2.4. As the first two components (PCs) were found to explain more than 98% of the variance, only these were selected for modeling streamflows. Next, the correlation values of different pairs (streamflow anomaly and two PCs) for different lags (1–3 months) were computed. PCs from predictor variables lagged by only 1 month were adopted for streamflow forecasting, as significant correlations were observed at this lag for both WS I and WS II.

Table 2.4 Principal components and the explained variance

Principal Component	Eigenvalues		Explained Variance (%)	
	WS I	WS II	WS I	WS II
1	4158.98	3535.89	80.52	81.17
2	943.94	773.00	18.27	17.75
3	33.95	29.50	0.66	0.68
4	22.83	11.54	0.44	0.26
5	2.96	3.13	0.06	0.07
6	2.19	2.59	0.04	0.06
7	0.51	0.57	0.01	0.01

2.5.2 Analysis of Asymmetric Archimedean Copula

The joint dependence between the streamflow anomaly, PC-1 and PC-2 requires that the nature of association between them be identified. The scatter plots of the pairs of predict and predictor variables indicated a higher degree of dependence between the streamflow anomaly and PC-1 with a correlation of 0.43 and 0.37 for WS I and WS II, respectively. The correlation between streamflow anomaly and PC-2 is 0.08 and 0.02, respectively, for WS I and WS II, whereas the first two PCs are uncorrelated by nature. Correlations between higher order PCs are very close to zero.

The scatter plots indicate that the pairs of variables have different bivariate dependence structures that cannot be modeled by the symmetric copulas (not included here for brevity). The Kendall's τ values of the various pairs of these variables are listed in Table 2.5. Given this nature of dependence, a class of asymmetric Archimedean copulas were adopted wherein the streamflow anomaly and PC-1 was coupled by a copula C_2, and this structure was then associated with PC-2 by another copula C_1.

From the streamflow anomaly values and the two PCs, their rank-based nonparametric marginal probabilities u_1, u_2, u_3, respectively, were calculated for modeling the copula function. The properties of asymmetric Archimedean copulas are mentioned in section 2.4.2. However, as the study data set did not conform to the requirement of the M12 nested 3-copula family that $\tau_{12}, \tau_{21}, \tau_{13} \in [0.333, 1]$ (Table 2.2), this copula family was rejected for both study watersheds.

2.5.3 Parameter Estimation

The parameters of the nested copula were estimated using the canonical maximum likelihood (CML) method [Genest et al., 1995; Kojadinovic and Yan, 2011]. The parameter values must conform to the range specified for each class of copula. The condition that the more nested variables have a stronger degree of dependence among them i.e. $\theta_2 \geq \theta_1 \in [0, \infty)$ was satisfied by the M3 and M4 families, and the condition $\theta_2 \geq \theta_1 \in [1, \infty)$ was satisfied by the M5 and M6 families of copula. The estimated values of the copula parameters and the maximum likelihood value obtained for each of the copula families are listed in Table 2.5.

2.5.4 Goodness-of-fit Tests

From the copula families evaluated in the study, the best copula was selected using popular goodness-of-fit measures. The probability distribution function of different copula families and the empirical copula are plotted in Figure 2.2. The performance statistics computed for the probability distribution function between the empirical and

estimated copulas are given in Table 2.6. The M6 copula family was found to have lowest value of S_n and T_n statistics calculated for WS I. The goodness-of-fit for this copula family is also evident from Figure 2.2(a). The lowest value of S_n and T_n was obtained for M6 copula in the case of WS II. It also provided the best distribution fit among all copula models in Figure 2.2b. Plots in Figure 2.3 show the performance of only the M6 copula for different months, suggesting that the dependence structure of the first two principal components of anomalies of the hydroclimatic variables and streamflow anomalies could be modeled by the same M6 copula family for all months in both the watersheds.

Table 2.5 Parameter θ for different copulas

Nested Copula Family	Maximum Likelihood Estimate					
	θ_1		θ_2		Maximum likelihood value	
	WS I	WS II	WS I	WS II	WS I	WS II
M3	0.005	0.185	3.35	2.71	55.71	47.12
M4	0.005	0.001	0.69	0.63	45.34	48.93
M5	1.08	1.10	1.57	1.35	44.73	28.33
M6	1.04	1.05	1.45	1.31	56.17	41.01

Table 2.6 Goodness-of-fit test statistics for different copulas

Nested Copula Family	S_n		T_n	
	WS I	WS II	WS I	WS II
M3	0.064	0.061	0.038	0.038
M4	0.105	0.116	0.051	0.044
M5	0.046	0.053	0.040	0.044
M6	0.043	0.043	0.038	0.041

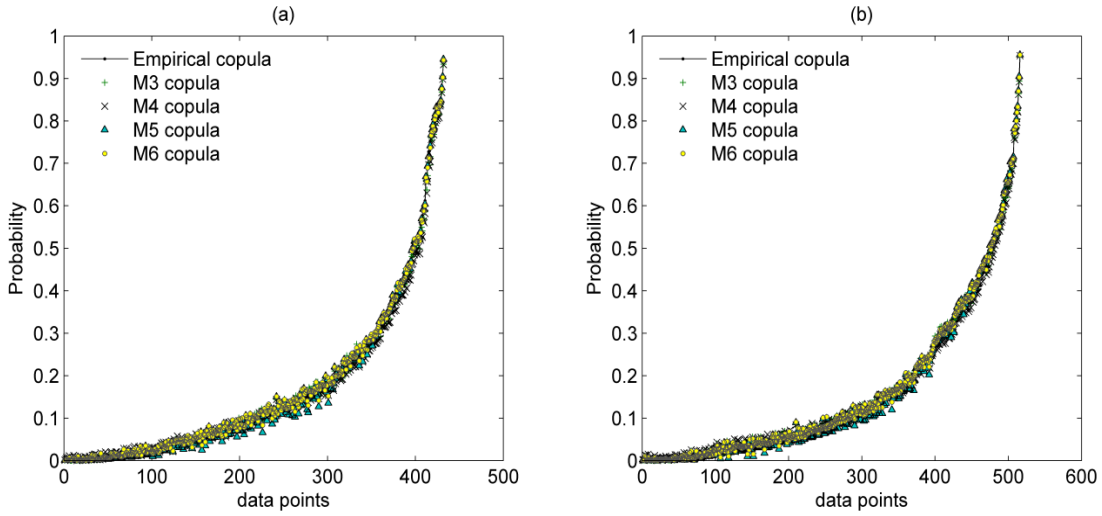

Figure 2.2 Comparison plots of probability distributions of different copula families used in (a) WS I and (b) WS II

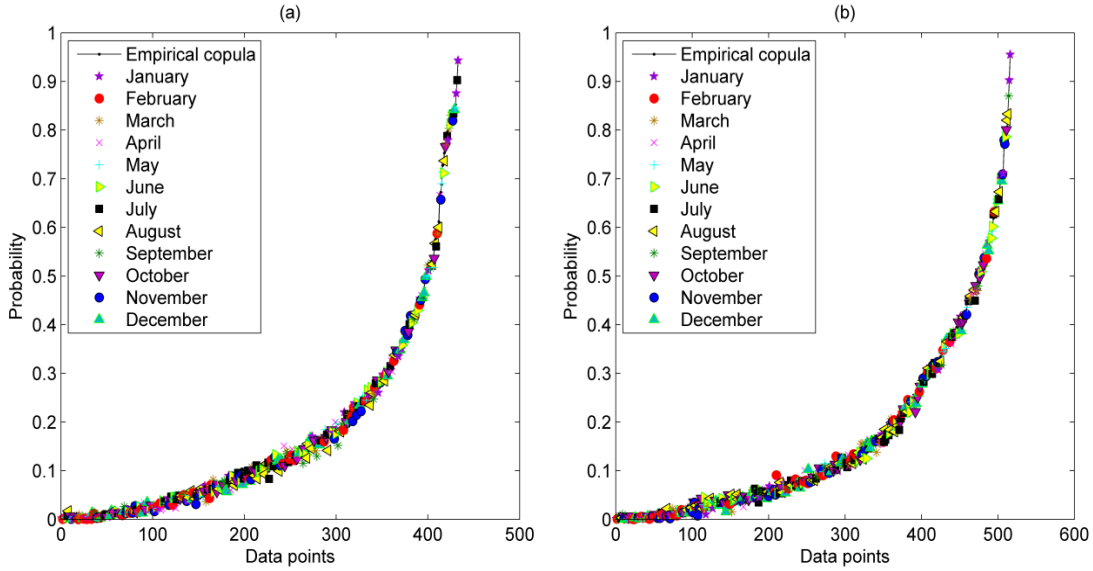

Figure 2.3 Plots showing M6 copula fit for each month in (a) WS I and (b) WS II

2.5.5 Streamflow Prediction Using Copula

Given u_2 and u_3 (the rank-based values of PCs extracted from the predictors), the probability distribution of u_1 (derived from streamflow anomalies) was generated using the M6 copula model (Table 2.2). The streamflow anomalies corresponding to different quantiles were calculated from this CDF. The rank-based non-parametric marginal probabilities at 0.025, 0.5 and 0.975 quantiles were calculated and transformed into the streamflow anomaly values; subsequently, the estimates of streamflows for the next month were obtained. Streamflows simulated for the model development period were compared with the observed flows for evaluating model performance.

The model developed for WS I was tested for the period January 1994 to December 2010, while model testing was carried out for the period 1991-2010 for WS II. The PCA coefficients obtained for predictors during model development period were used to obtain the PCs for the testing period as well. The predicted streamflow values for the model development and testing periods are compared with corresponding observed flows in Figures 2.4a, 2.4b, 2.5a and 2.5b for the two watersheds. The uncertainty in the predictions is quantified by the plot of interquantile range of predicted streamflows. Most of the observed flows lie within the predicted range during the model development periods in WS I. Typically, low flows in the late 1960s and 1970s are in close agreement with the expected values of streamflows obtained from the model (Figure 2.4a). The low flows during the testing period, especially in the 1990s, match well with the expected values in Figure 2.4b. However, this is not the case with high flows in WS I during both training and testing periods, where 1 month lead forecasts underestimate the observed

peaks. In WS II, the recorded flows fall within the range of probabilistic predictions offered by the developed model. In Figure 2.5a, the predicted low flows in the 1950s, 1960s, and 1980s conform to observations. During the testing period also, the model performed well with low flow predictions (Figure 2.5b). The peak flows for both training and testing periods were typically underestimated perhaps because of the small numbers of training samples in this range. Additionally, the box plots for model development and testing periods in WS I and WS II in Figures 2.4c and 2.5c, respectively, indicate that though the model performance is not satisfactory in the case of high flows, low flows are estimated well. Overall, the predictive capability of the model was found to favor low flow conditions, prompting us to explore the development of droughts over the two study watersheds. The coefficients of determination (R^2) values obtained were 0.64 and 0.53, respectively, for the model development and testing periods in WS I, and 0.58 and 0.50, respectively, for WS II. Comparisons with state-of-the-art statistical models [Tripathi and Govindaraju, 2008] using the same set of predictors for streamflow showed similar performance, but the results are not reported here for brevity.

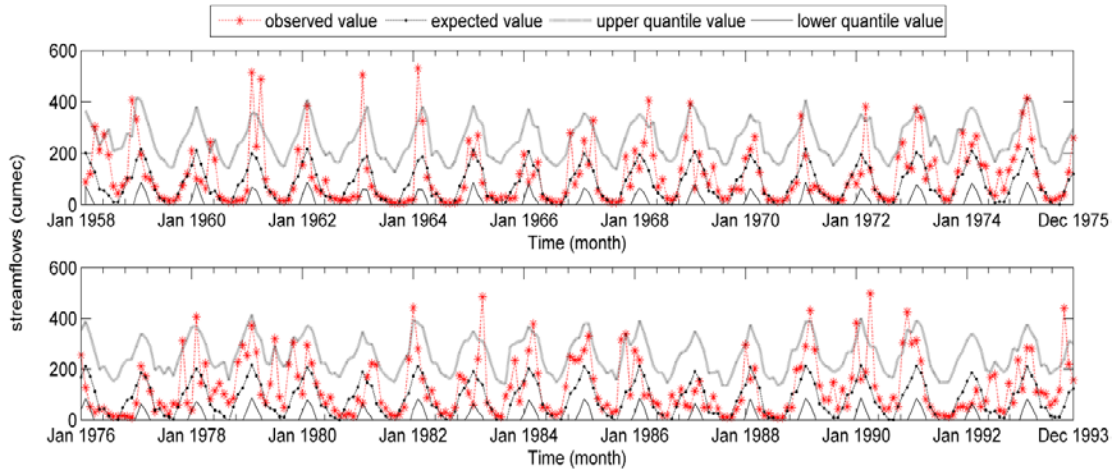

Figure 2.4a Comparison plots of observed and predicted streamflows in WS I during model development period (lower and upper quantile curves correspond to 0.025 and 0.975 quantiles, respectively)

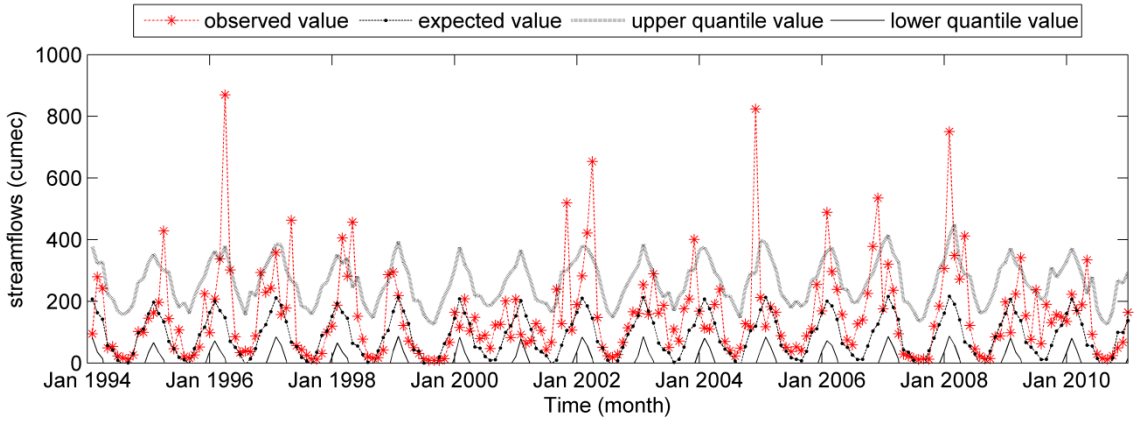

Figure 2.4b Comparison plots of observed and predicted streamflows in WS I during model testing period (lower and upper quantile curves correspond to 0.025 and 0.975 quantiles, respectively)

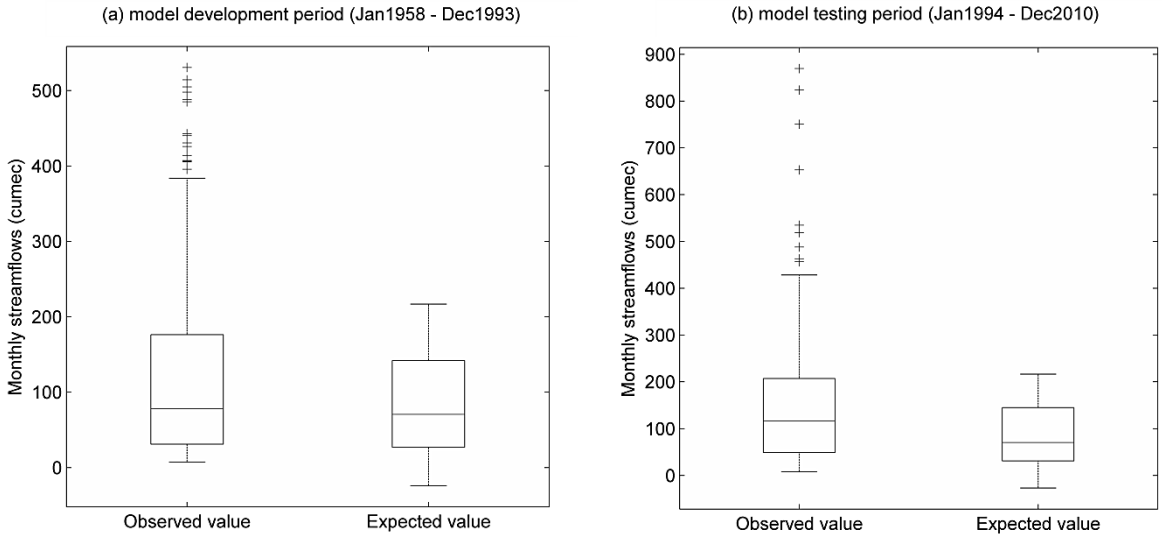

Figure 2.4c Box plots for observed and predicted (expected) values of monthly streamflows during model development and testing periods in WS I. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted with a '+' symbol

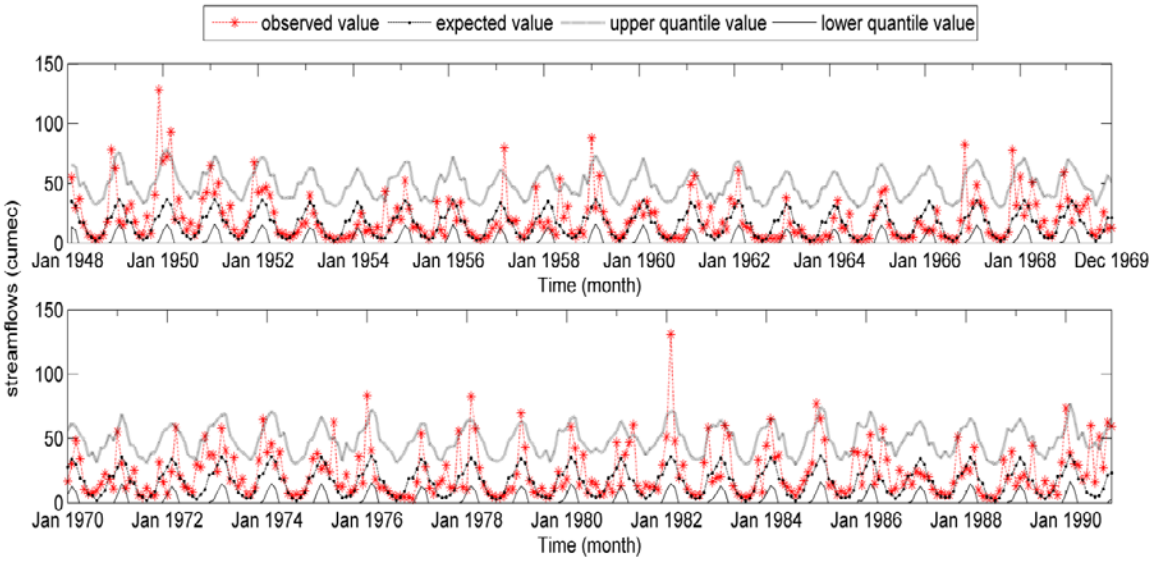

Figure 2.5a Comparison plots of observed and predicted streamflows in WS II during model development period (lower and upper quantile curves correspond to 0.025 and 0.975 quantiles, respectively)

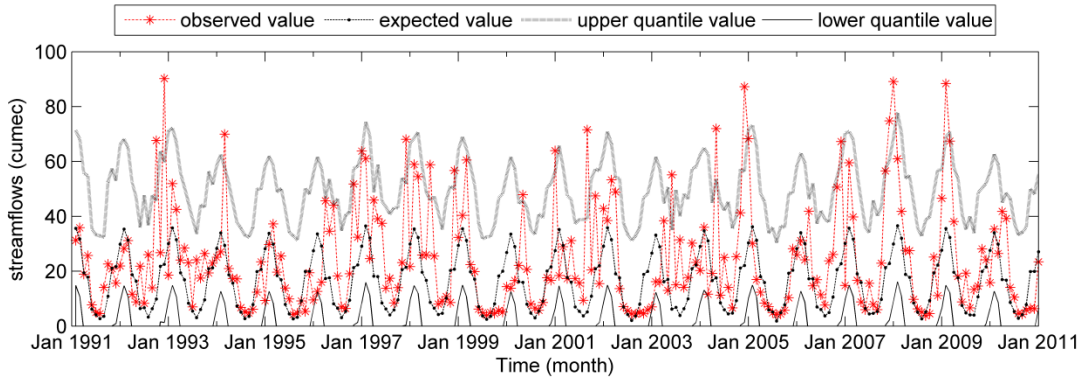

Figure 2.5b Comparison plots of observed and predicted streamflows in WS II during model testing period (lower and upper quantile curves correspond to 0.025 and 0.975 quantiles, respectively)

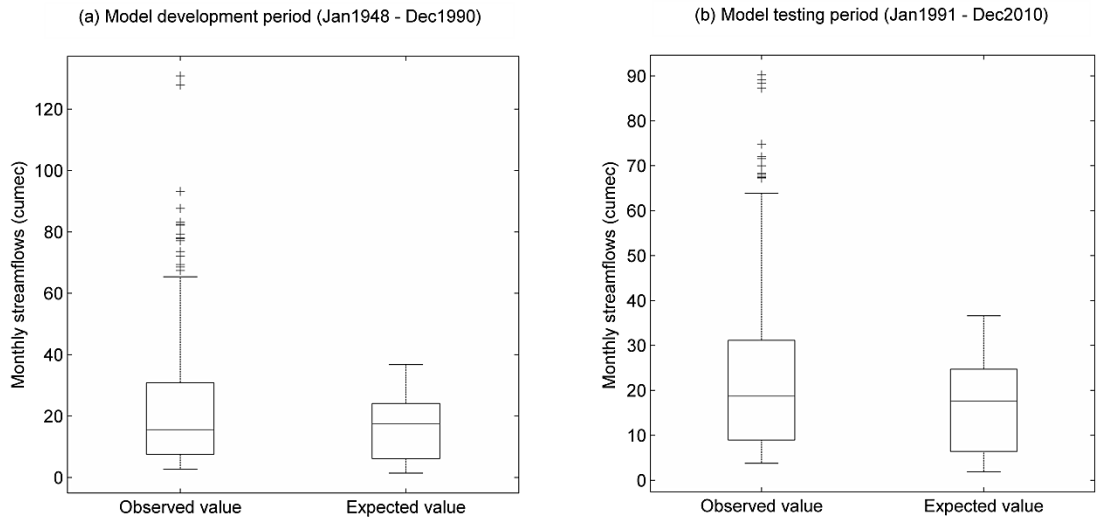

Figure 2.5c: Box plots for observed and predicted (expected) values of monthly streamflows during model development and testing periods in WS II. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted with a '+' symbol

2.5.6 Drought Analysis

The results of the drought analysis carried out for the model development period (January 1948-December 1993) for WS I are shown in Figure 2.6a. There were few occurences of D3 and D4 classes of droughts during the model development periods, and mild (D0) and moderate (D1) droughts prevailed in most of the drought months. The drought index values obtained from the expected streamflows provided good forecasts of dry as well as wet conditions. The drought analysis was then carried out for the testing period and compared with the observed conditions. Few occurences of D2 and D1 classes of droughts marked the testing period. Wet conditions dominated during this period, with most of them being underestimated by the model (Figure 2.6b). The plots for drought indices calculated for WS II in Figure 2.7a and 2.7b also indicate that different drought categories were better predicted than the wet categories. The sequences of drought months in different sub-periods during the entire model development and testing periods were also well predicted.

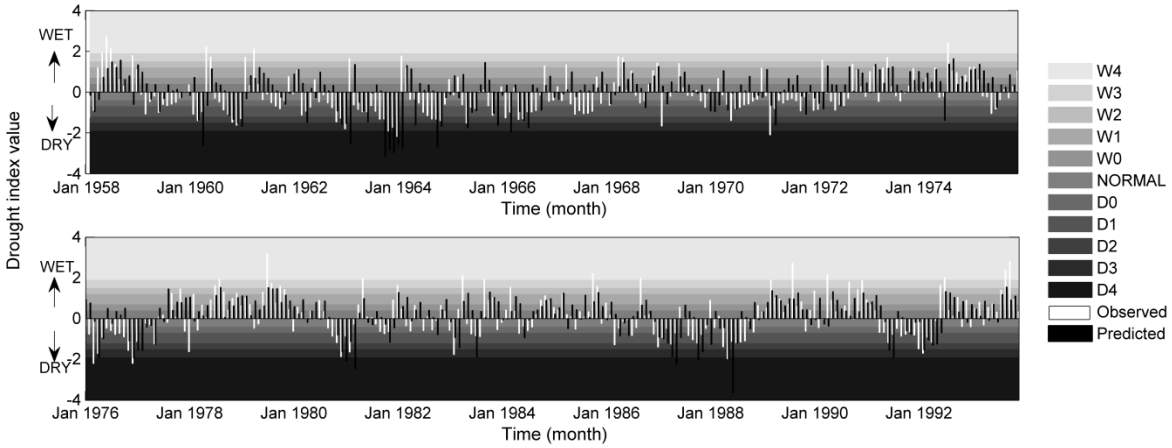

Figure 2.6a Drought index values during the model development period in WS I

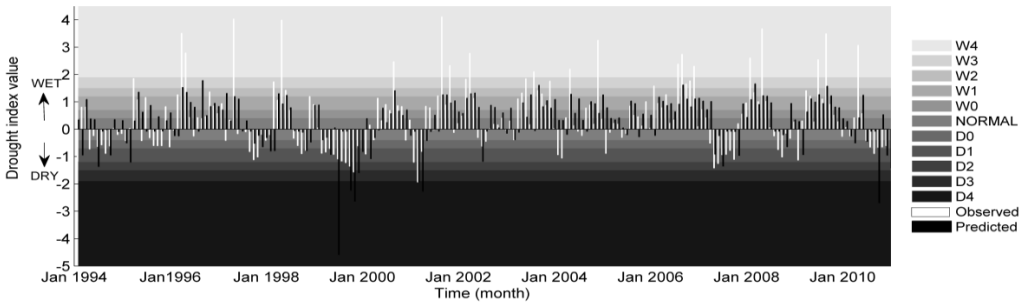

Figure 2.6b Drought index values during the model testing period in WS I

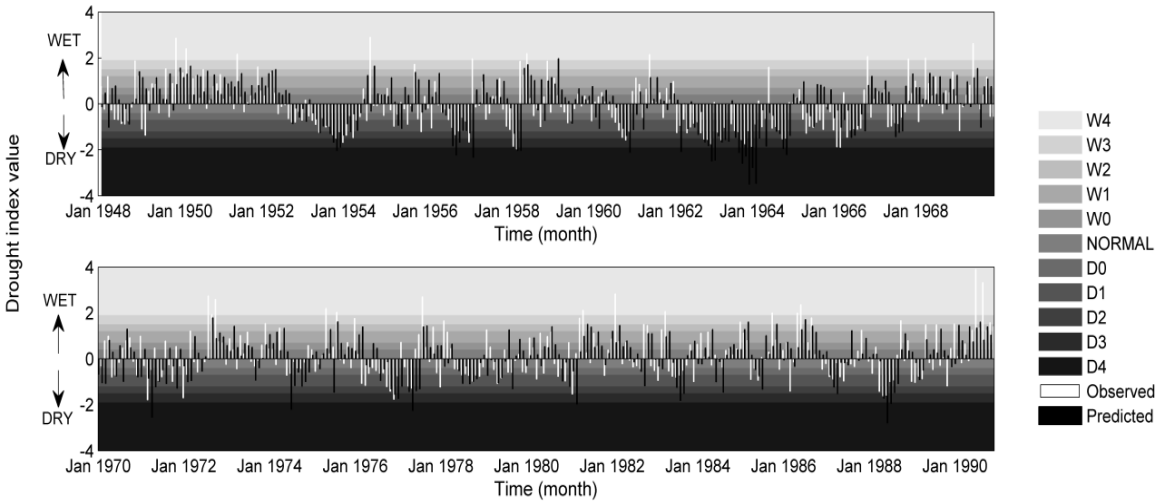

Figure 2.7a Drought index values during the model development period in WS II

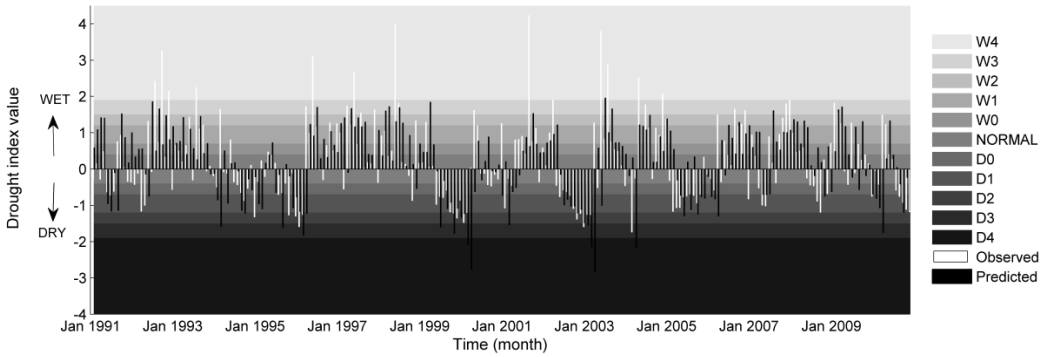

Figure 2.7b Drought index values during the model testing period in WS II

Apart from visual inspection, the model performance for multiple category classification of streamflows was assessed by computing the contingency coefficient C, proposed by Pearson [1904]. This coefficent is a measure of degree of association between multiple categories in a contingency table classifying N samples [Gibbons and Chakraborti, 2011] and mathematically expressed as:

$$C = \sqrt{\left(\frac{Q}{Q+N} \right)} \tag{2.12}$$

where, Q is a statistic that tests the null hypothesis that there is no association between observed and predicted categories.

Q is expressed as:

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{\left(NX_{ij} - X_{i.}Y_{.j} \right)^2}{NX_{i.}Y_{.j}} \tag{2.13}$$

where r and k are the number of categories, X_{ij} is the number of cases falling in i^{th} observed and j^{th} predicted category, $X_{i.} = \sum_{j=1}^{k} X_{ij}$ and $Y_{j.} = \sum_{i=1}^{r} X_{ij}$.

The statistic Q approximately follows chi-square distribution with degrees of freedom (dof) equal to $(r-1)(k-1)$. Thus, the null hypothesis (no association) can be rejected if the p-value is very low. Higher values of C correspond to better association. The value of C cannot exceed 1 theoretically and has an upper bound of $C_{\max} \left(= \sqrt{(t-1)/t} \text{ where } t = \min(r,k) \right)$ [Gibbons and Chakraborti, 2011]. The ratio $C / C_{\max}$ is often used as a measure of degree of association.

In order to ensure sufficient data for robust statistics, a contingency table with three different categories: dry, normal and wet was prepared. The extreme categories were merged to ensure that the observations and predictions are available sufficiently in all categories. These contingency tables are shown in Tables 2.7a and 2.7b for WS I and WS II, respectively. Thus, both r and k are 3, and dof is 4. The statistic Q, contingency coefficient C, and the measure of degree of association $C/C_{\max}$ are shown at the end of Tables 2.7a and 2.7b. The low p-values for the statistic Q indicate that the null hypothesis of no association between observed and predicted categories should be rejected. The degree of association was found to be reasonable for both the watersheds during model development as well as testing periods.

Table 2.7a Contingency table and degree of association between observed and predicted drought categories for WS I

Predicted Category	Model development period (1958-1993)			Model testing period (1994-2010)		
	Observed Category			Observed Category		
	Dry	Normal	Wet	Dry	Normal	Wet
Dry	71	18	11	18	9	5
Normal	78	38	39	30	23	16
Wet	31	62	84	8	30	65
Q		85.75			52.92	
DOF		4			4	
p-value		< 0.0001			< 0.0001	
C		0.407			0.454	
C_{max}		0.817			0.817	
C/C_{max}		0.498			0.556	

Table 2.7b Contingency table and degree of association between observed and predicted categories for WS II

Predicted Category	Model development period (1948-1990)			Model testing period (1991-2010)		
	Observed Category			Observed Category		
	Dry	Normal	Wet	Dry	Normal	Wet
Dry	107	24	23	40	14	15
Normal	63	50	45	24	19	13
Wet	33	80	91	12	39	64
Q		105.54			54	
DOF		4			4	
p-value		< 0.0001			< 0.0001	
C		0.412			0.429	
C_{max}		0.817			0.817	
C/C_{max}		0.505			0.525	

2.5.7 Extraction of Drought Triggers

Using the modeled asymmetric copula dependence function, the conditions that trigger hydrological droughts or extremes in the watershed were examined. The triggers for various streamflow conditions were generated using the conditional copula. The procedure is illustrated as follows. Given a certain streamflow anomaly quantile α, let y_1^α and y_2^α correspond to the first and second PCs conditioned on the streamflow anomaly value. The quantities y_1^α and y_2^α are obtained from the M6 copula for the particular watershed. Since these two PCs explain over 98% of the total variation, the other principal components remain unaffected by the choice of the streamflow quantile. Our goal is to find the expected values of the precursor variables $x_i^\alpha, i = 1, 2, ..., 7$ that

would correspond to this particular streamflow quantile. If a_{ij} are the PCA coefficients for the data set, then the following equation provides the conditional expectation of the precursor variables:

$$[A]\{x^{\alpha}\} = \{y^{\alpha}\} \tag{2.14}$$

where a_{ij} is the ij^{th} element of the matrix $[A]$, y_1^{α} and y_2^{α} are computed from the M6 copula, and y_j, $j = 3,4,...,7$ are simply the expected values of the principal components (≈ 0).

The expected values of PC-1 and PC-2 conditioned on various streamflow anomaly quantiles(corresponding to different α values) are shown in Table 2.8 for both watersheds. The expected anomaly values of all the predictor variables for different values corresponding to different streamflow anomalies are shown in Tables 2.9a and 2.9b. Low flows correspond to smaller values of soil moisture, temperature, precipitation, evaporation and runoff of the previous month in both watersheds. Sea level pressure anomaly varied inversely with the streamflow anomaly for WS I and WS II, suggesting that increase in sea level pressure from the long term mean can enhance the chances of droughts in the regions. Increase in wind speed was found to trigger droughts in WS I, in contrast to the trend observed in the case of WS II. The dissimilar trends in some variables suggest that drought triggers are likely to be specific to each watershed.

The conditional expectations of anomalies of different precursors corresponding to different streamflow quantiles (Table 2.9) were utilized to develop potential triggers for each drought category. The long-term monthly means of hydroclimatic variables were added to their expected anomaly values to carry out this analysis. The resulting precursor

values were then associated with the 1 month lead drought index values. From the expected streamflow anomaly, streamflows for each month were computed and corresponding drought indices were calculated. The trigger analysis is limited to low flow conditions corresponding to drought reflecting the better model performance for flows in this range. The plots in Figures 2.8a and 2.8b show the expected precursor range in each month obtained for different drought classes for WS I and WS II, respectively. If the values of the hydroclimatic variables fall within the suggested range for any class of drought, then that drought would likely occur in the succeeding month. For WS I, soil moisture, precipitation, and runoff are able to offer a range of predictor values for different drought categories as shown in Figure 2.8a. Some months (May to July) do not show any range of potential predictor values for certain drought classes, implying the likelihood of such droughts being very low in those periods in WS I. While soil moisture, precipitation, and runoff show some variability with drought classes in WS II, the other variables stay within a very tight band for any given month (Figure 2.8b). Thus, only these three variables are capable of resolving amongst different drought classes for the study watersheds. Low variability is manifested in the expected anomaly values of temperature, evaporation, sea-level pressure, and wind speed in Table 2.9.

The precursor ranges developed in this manner were validated by means of scatter plots between the observed and modeled values of variables over the model development and testing periods (Figures 2.9a and 2.9b) for all classes of droughts. These scatter plots demonstrate good agreement between the observed and modeled triggers in both watersheds. The scatter is less in the case of soil moisture, precipitation, runoff, evaporation, and temperature in both watersheds. Among the predictors, wind speed

shows the most scatter making it the least reliable precursor for both watersheds. The modeled triggers for soil moisture, precipitation, and runoff values are underpredicted compared to observations during calibration as well as validation. Additionally, correlation values for all the trigger variables were calculated and tabulated in Table 2.10. High correlations in some predictors (for example, temperature and evaporation in WS I and WS II), however, were not useful as they were found incapable of resolving among the different drought categories.

The results indicate that drought trigger information retrieved in this manner has potential for applications in hydrologic drought preparedness. Even though individual variables show scatter, if multiple variables fall close to their trigger values, the confidence in their effectiveness as hydrologic drought triggers will improve. Hence, the combined behavior of predictor variables needs to be considered when estimating potential drought triggers.

Table 2.8 Expected principal component values for various quantiles of streamflow

Streamflow anomaly quantile	Streamflow anomaly (cumecs)		Expected PC-1 value		Expected PC-2 value	
	WS I	WS II	WS I	WS II	WS I	WS II
0.01	-172.84	-27.65	-49.10	-34.41	-3.79	-5.33
0.1	-99.47	-17.27	-34.31	-23.69	-3.25	-4.86
0.2	-63.19	-10.62	-26.54	-17.36	-2.44	-4.03
0.4	-27.13	-4.86	-15.73	-6.11	-1.97	-3.26
0.5	-16.94	-3.26	-8.79	-1.98	-1.64	-3.10
0.6	-5.71	-1.20	-1.63	2.98	-0.96	-1.79
0.8	58.38	8.79	25.20	18.02	0.57	-0.86
0.9	123.30	20.38	43.28	31.62	1.08	0.54
0.99	310.66	53.54	121.66	80.46	4.20	5.56

Table 2.9 Conditional expectations (in terms of anomalies of hydro-climatic variables) associated with streamflow anomaly

Expected Streamflow anomaly (cumecs)	(a) Hydro-climatic triggers in terms of expected values of anomalies in WS I						
	Soil moisture anomaly (mm)	Temperature anomaly (°C)	Precipitation anomaly (mm)	Evaporation anomaly (mm)	Sea level pressure anomaly (mbar)	Wind speed anomaly (m/s)	Runoff anomaly (mm)
-172.84	-37.33	-0.0032	-31.28	-1.26	0.21	0.0017	-7.18
-99.47	-25.74	-0.0072	-22.35	-0.86	0.15	0.0005	-5.02
-63.19	-19.94	-0.0050	-17.23	-0.67	0.12	0.0005	-3.88
-27.13	-11.52	-0.0072	-10.63	-0.38	0.07	-0.0003	-2.31
-16.94	-6.12	-0.0085	-6.38	-0.19	0.05	-0.0008	-1.30
-5.71	-0.74	-0.0070	-1.72	-0.01	0.01	-0.0008	-0.25
58.38	19.96	-0.0096	14.94	0.69	-0.10	-0.0024	3.66
123.30	34.22	-0.0157	25.74	1.18	-0.17	-0.0040	6.29
310.66	95.51	-0.0345	73.30	3.29	-0.49	-0.0100	17.70
Expected Streamflow anomaly (cumecs)	(b) Hydro-climatic triggers in terms of expected values of anomalies in WS II						
	Soil moisture anomaly (mm)	Temperature anomaly (°C)	Precipitation anomaly (mm)	Evaporation anomaly (mm)	Sea level pressure anomaly (mbar)	Wind speed anomaly (m/s)	Runoff anomaly (mm)
-27.65	-27.61	-0.127	-20.92	-0.94	0.126	-0.046	-3.42
-17.27	-18.44	-0.102	-15.45	-0.62	0.097	-0.036	-2.37
-10.62	-13.29	-0.080	-11.73	-0.44	0.075	-0.028	-1.74
-4.86	-3.80	-0.050	-5.75	-0.17	0.042	-0.016	-0.64
-3.26	-0.27	-0.041	-3.66	0.01	0.031	-0.012	-0.24
-1.20	3.47	-0.016	-0.17	0.13	0.008	-0.004	0.26
8.79	16.20	0.023	7.75	0.57	-0.036	0.012	1.74
20.38	27.44	0.065	15.39	0.95	-0.079	0.027	3.08
53.54	67.84	0.215	42.84	2.34	-0.238	0.082	7.89

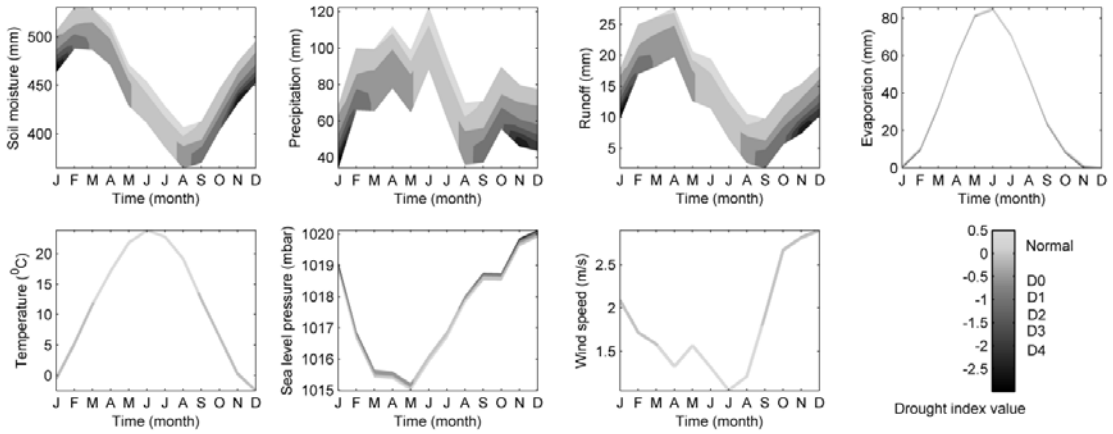

Figure 2.8a Contour plots showing expected ranges of different hydro-climatic variables as precursors to droughts in WS I

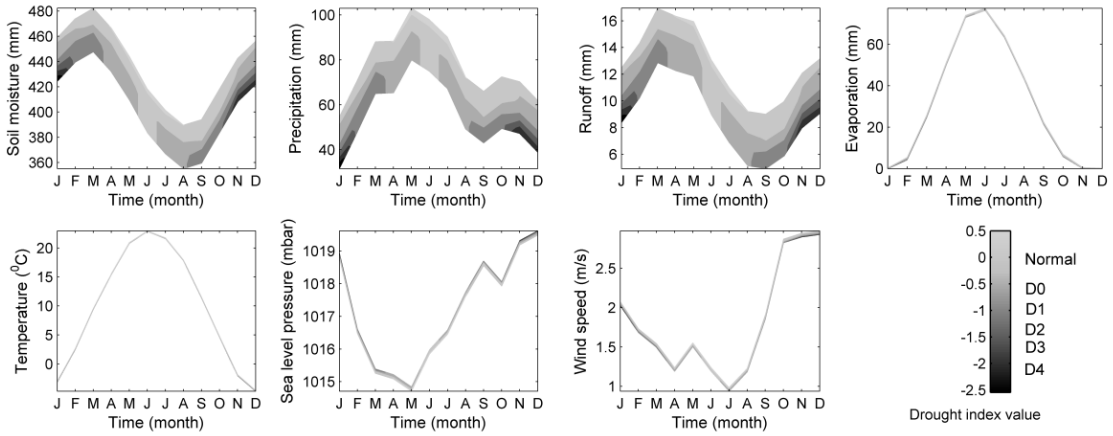

Figure 2.8b Contour plots showing expected ranges of different hydro-climatic variables as precursors to droughts in WS II

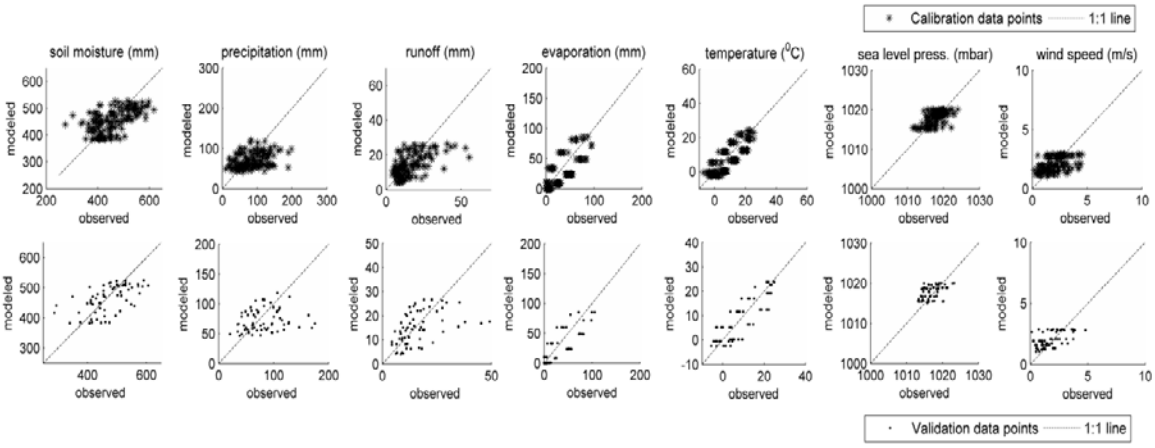

Figure 2.9a Scatter plots of different hydro-climatic precursors (modeled versus observed) for model development and testing periods in WS I

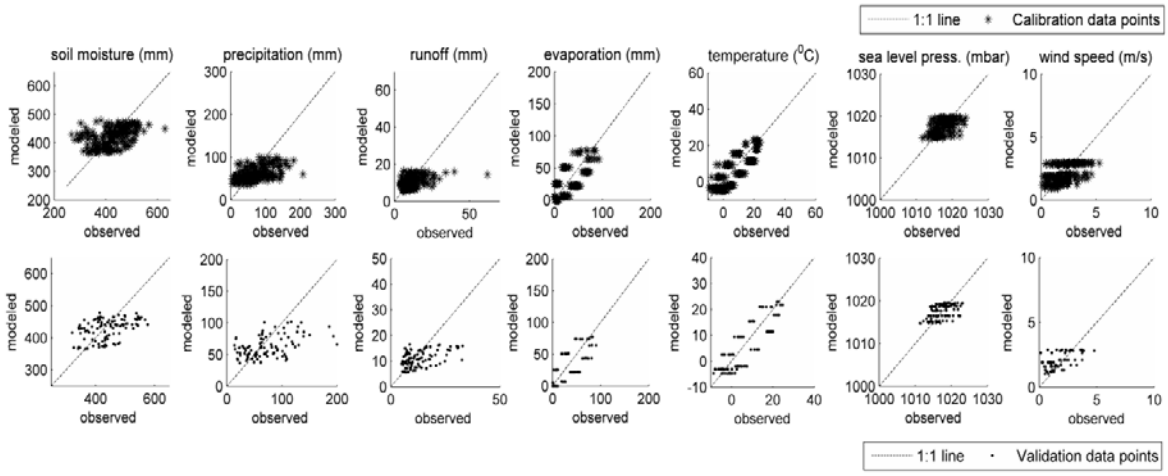

Figure 2.9b Scatter plots of different hydro-climatic precursors (modeled versus observed) for model development and testing periods in WS II

Table 2.10 Correlation values between observed and modeled drought precursors.

Hydro-climatic precursor	WS I		WS II	
	calibration	validation	calibration	validation
Soil moisture	0.57	0.58	0.41	0.44
Precipitation	0.35	0.29	0.48	0.44
Runoff	0.59	0.47	0.41	0.45
Evaporation	0.80	0.83	0.80	0.82
Temperature	0.81	0.82	0.82	0.85
Sea-level pressure	0.58	0.43	0.50	0.52
Wind speed	0.45	0.56	0.48	0.52

2.6 Summary and Conclusions

This chapter provides a novel method for developing drought triggers by combining the strengths of PCA for dimensionality reduction and copulas for modeling the joint dependence between variables. The first two PCs were found capable of explaining the variability in the anomaly set of predictor variables for both study watersheds. The joint dependence of the streamflow anomaly and the two principal components was modeled by a scale-free association using a suitable asymmetric 3-copula selected based on goodness-of-fit statistics. The developed model was first tested for forecasting streamflows in two study watersheds.

The chapter focused on 1-month lead predictions because correlations between the principal components and streamflow anomaly diminished rapidly beyond a lag of 1 month. Under-prediction of peak flows was observed in the results of both watersheds, but low streamflows were reasonably predicted allowing hydrologic drought studies. Drought index values based on standardized flows were computed to identify the occurrences of droughts during the model development and testing periods in the two study regions.

The conditional dependence of the principal components PC-1 and PC-2 on streamflow anomaly was used to determine the drought triggers in the two watersheds. The precursors to droughts were expressed in terms of the anomaly values of the climatic variables. Negative anomalies of soil moisture, precipitation, evaporation, temperature, and runoff, and increased sea-level pressure and wind speeds were obtained as potential drought triggers for WS I.

Similarly, increased sea level pressure conditions and reduced soil moisture, precipitation, evaporation, temperature, runoff, and wind speeds from their respective long-term means led to drought conditions in WS II.

Further, the patterns of various hydroclimatic variables as potential precursors to different categories of droughts were examined for the two watersheds. The ranges of predictor values that led to different drought conditions were estimated from the expected precursor values for low streamflow quantiles. The trigger analysis results were validated by comparing the observed hydroclimatic variables with their expected trigger values for the model development and testing periods. The correlation values computed indicated that the analysis could yield reliable information on the pattern of drought triggers for both the watersheds.

The following conclusions are derived:

i. Drought triggers are likely to be specific to watersheds. Even though the two study watersheds are located in the same part of the world and have similar land use distribution, local conditions influence streamflows especially at monthly time scales.

ii. Using copulas, conditional expectations of first two PCs based on different quantiles of streamflow anomalies provide a method for estimating drought triggers. Among all the precursors, soil moisture, precipitation, and runoff showed the greatest potential for assessing different classes of droughts for both watersheds. The other variables, despite showing strong seasonal trends, demonstrated little capability for resolving the different drought classes.

iii. Validation results for triggers over all drought classes show results with different degrees of variability. Even with the scatter present for single (individual) variables, if triggers for multiple variables fall within expected ranges, the confidence in the trigger would improve. Hence, it is recommended that precursors for droughts be examined in combination by using multiple input variables.

Even though the results and conclusions are specific to study watersheds, the method shows promise for application to different watersheds. An important limitation is that the level of dimensionality reduction that can be achieved in different watersheds cannot be known a priori. If multiple predictors were to be important, the model for constructing the joint distribution would be too complex for practical purposes except in limited cases modeled using Gaussian copulas. Data limitations also continue to be a serious challenge for many hydrologic studies. Large amount of data need to be used for capturing the trigger behaviors in drought studies. The model development and testing periods were short in this study, and the methodology performs reasonably well even for the small record lengths available here. Future efforts employing more hydroclimatic variables and different watersheds will help develop better understanding of trigger mechanisms for droughts.

CHAPTER 3. PREDICTOR SELECTION FOR STREAMFLOWS USING A GRAPHICAL MODELING APPROACH

3.1 Abstract

Streamflows are influenced by various hydroclimatic variables in complex ways. Accurate prediction of monthly streamflows requires a clear understanding of the dependence patterns among these influencing variables and streamflows. A graphical modeling technique, employing conditional independence, is adopted in this study to quantify the interrelationships between streamflows and a suite of available hydroclimatic variables, and to identify a reduced set of relevant variables for parsimonious model development. The nodes in the undirected graph represent relevant variables, and the strengths of the connections among the variables are learnt from the data. The graphical modeling approach is compared to the state-of-the-art method for predictor selection based on partial mutual information. For a synthetic benchmark dataset and a watershed in southern Indiana, USA, the graphical modeling approach shows more discriminating results while being computationally efficient. Along with artificial neural networks and time series models, results of the graphical model are used for formulating a variational relevance vector machine to predict monthly streamflows and perform probabilistic classification of hydrologic droughts in the watershed being studied. The parsimonious models developed for prediction at different lead times performed as well as the non-

parsimonious models during both the calibration and testing periods. Drought forecasting for the study watershed at 1-month lead time was performed using the two selected predictors−soil moisture and precipitation anomalies alone, and the model performance was evaluated. The graphical model shows promise as a tool for predictor selection, and for aiding parsimonious model development applications in statistical hydrology.

3.2 Introduction

Prediction of streamflows is an important component of hydrologic modeling, water quality, resource management and impact assessment studies. The utility of models in analysis and design of water resources systems is well known, be it for understanding the processes involved, to simulate system behavior and interactions, and to perform risk analysis [Praskievicz and Chang, 2009]. Hydrologists and water engineers around the globe have built robust prediction and forecasting models, yet there are several uncertainties associated with spatial and temporal variability in hydrological variables replicated in these models [Tian et al., 2014; Livneh and Lettenmaier, 2012]. Unplanned excess or shortage of water supply affects the socio-economic status of dependent areas through floods or droughts [Alcamo et al., 2007; Burn et al., 2008].

Monthly streamflow prediction at a basin scale is a challenging problem because of the complex roles of multiple interacting hydro-climatic variables such as precipitation, temperature, soil moisture, mean sea level pressure, sea surface temperature, runoff, wind speed and mean sea level pressure, that contribute to flow generation. Thus, while streamflows are known to depend on various hydroclimatic

variables, dependence patterns among these predictor variables and streamflows are site-specific, and methods for identifying relevant predictor variables are needed for forecasting purposes.

Competent predictor selection is an important part of the development of skillful forecast models [Makkeasorn et al., 2008], and poses a challenge for streamflow prediction models. Apart from selecting variables based on our understanding of the physical system [Robertson and Wang, 2009], temporal relations between the predictor set and predictand need to be accounted for using techniques such as time series correlation and cross-correlation analyses [Besaw et al., 2010]. Inclusion of all possible hydroclimatic variables that govern streamflows at a basin scale, and at multiple lags, will yield a prohibitively large number of variables in the predictor set resulting in highly complex prediction models and pose serious challenges in parameter estimation, in addition to being burdened with redundancy.

Prioritizing the relevant features in the vast set of potential predictor variables has several advantages: (i) better understanding of the data, (ii) improvement in classification of extremes, and (iii) avoiding the *curse of dimensionality*. Feature transform techniques (principal components analysis, PCA, and independent component analysis, ICA), and feature selection algorithms (wrapper, filter and online methods) have been used in several classification and pattern recognition studies [Maity et al., 2013; Maier et al., 2010; Crone and Kourentzes, 2010; Peng et al., 2005; Hsu et al., 2002]. Wrapper approaches utilize the performance of the resulting model to select the relevant features, whereas online methods incrementally add/remove variables during model development [Bonev, 2010]. Filter approaches, on the other hand, perform statistical tests on the

variable set, and extract input features possessing maximum mutual information with the desired output. In this regard, Sharma et al. [2000] used partial mutual information (PMI) to identify predictors of quarterly rainfall from a suite of hydroclimatic variables, and Hejazi and Cai [2009] employed minimum redundancy maximum relevance (MRMR) algorithm based on mutual information for input variable selection in a reservoir release prediction model. The PMI criterion facilitated selection of predictors by considering the partial or additional dependence added by a new variable to an existing predictor set. Bowden et al. [2005] investigated utility of two approaches: PMI in conjunction with general regression neural network (GRNN), and self-organizing map (SOM) with hybrid genetic algorithm (GA)-GRNN for input selection. Based on tests on synthetic data sets whose dependence relations are known, PMI-based method selected all significant inputs unlike the SOM-GAGRNN method that required an appropriate objective function and involved additional parameters (population size and number of generations) of the genetic algorithm. A major drawback in applying the PMI algorithm to large data sets is the computational burden in computing the 95^{th} percentile randomized sample statistic [May et al., 2008]. Modifications were made to the PMI algorithm-based predictor selection by May et al. [2008] and Fernando et al. [2009], using a Hampel distance-based score [Davies and Gather, 1993] as the termination criterion. Besides, there are ranking measures for variables based on information theory such as Shannon entropy, Kullback-Leibler measure, Euclidian distance, and Kolmogorov dependence that are commonly used in machine learning [Bonev, 2010]. Several hydrologic studies have used correlations and partial correlations between the predictors and predictand, in an iterative fashion, to extract the most useful predictors [Phatak et al., 2011; Traveria et al., 2010;

Prasad et al., 2010]. In addition to these, Gamma test (GT), and forward selection (FS) are other popular techniques employed to reduce the dimensionality of an input variable set [Noori et al., 2011; Moghaddamnia et al., 2009]. Maity and Kashid [2011] developed a Birnbaum importance measure-based technique to identify the set of important inputs from an initial pool of predictor variables. A tree-based iterative input variable selection (IIS) scheme was recently proposed by Galelli and Castelletti [2013], yielding a rapid predictor selection algorithm. However, the sensitivity of this model to parameters requires trial and error based fine-tuning for the regression problem.

When multiple predictors are likely to govern the response of hydrological systems, probabilistic graphical models offer an attractive model-free method (i.e. by avoiding model performance assessment) for parsimonious predictor selection. A graphical model is a family of probability density functions that incorporate a specific set of conditional independence constraints listed in an independence graph [Jordan, 2004; Jensen and Nielsen, 2007; Whittaker, 2009]. A graph can therefore be perceived as a compact representation of interdependencies that exist in a multivariate distribution as well as a skeleton for factorizing a distribution. Establishing a graphical model is a powerful way of summarizing the interactions manifest within a set of variables. The technique offers (i) simplicity in condensing the multivariate data set without eliminating or obscuring any interesting associations, (ii) an ability to quantify the interrelationships between several variables by utilizing conditional independencies among variables, and (iii) an intuitive framework for statistical analysis of continuous data summarized by a correlation matrix [Lauritzen, 1996; Whittaker, 2009]. Graphical models are useful for describing and understanding many natural phenomena [Fiori et al., 2012]. Multi-scale

graphical models were used in climate dynamics to capture the interactions among Gaussian random variables in satellite imagery [Willsky, 2002] and to model spatial and temporal patterns of rainfall observed at multiple stations [Ihler et al., 2007]. Yu et al. [2012] proposed a copula Gaussian graphical model to capture the conditional dependence among extreme events across space, which could then be used to predict extreme values at unmonitored sites.

Once the predictor set has been identified, prediction models for streamflows can be built from the selected hydroclimatic variables using state-of-the-art regression techniques. Linear regression, artificial neural networks (ANNs), and autoregressive moving average models (ARMA) are popular approaches [Bowden et al., 2005; Wang et al., 2009; Gao et al., 2010]. Kernel-based approaches such as support vector machines (SVMs) and relevance vector machines (RVMs) have found several applications in hydrologic studies, and yield good predictions [Khalil et al., 2005; Asefa et al., 2006; Tripathi et al., 2006; Ghosh and Mujumdar, 2008; Karamouz et al., 2009; Dogan et al., 2009; Maity et al., 2010; Tripathi and Govindaraju, 2007, 2011; Kisi and Cimen, 2011; Hoque et al., 2012]. Variational RVMs [VRVMs; Bishop and Tipping, 2000; Faul and Tipping, 2001], for instance, operate in a fully Bayesian paradigm to deal with outliers that otherwise affect model robustness.

The main objective of this chapter is to propose graphical models as a novel approach to predictor selection for monthly streamflow prediction. The conditional independence structure between the predictand variable and predictors is extracted using a Gaussian graphical modeling technique to find the relevant predictors, and then this reduced variable set is utilized for streamflow prediction. The proposed method of

identifying predictor variables is shown to be superior to state-of-the-art methods. Such a graphical modeling-based approach for supervised predictor selection from a pool of interdependent hydroclimatic variables has not been evaluated in hydrologic applications. Following predictor selection, monthly streamflows for different lead times in future up to four months are forecasted using the reduced set of predictors at current time step and three statistical models, namely artificial neural networks (ANNs), autoregressive moving average model with exogenous inputs (ARMAX), and variational relevance vector machines (VRVM), to demonstrate the robustness of the predictor selection method across a suite of models. The application of this method is demonstrated for probabilistic classification of hydrologic droughts at monthly time step over a watershed in Indiana. The remainder of this chapter is organized as follows. In section 3.3, details are provided for the study area and data used for the present analysis. Section 3.4 describes the graphical model-based predictor selection methodology and its application to test cases and to future streamflows over the study area, followed by results and discussion in section 3.5. Summary and conclusions derived from the study are presented in section 3.6.

3.3 Study Area and Data Used

3.3.1 Study Area

The study was carried out over an agricultural watershed in southern Indiana, USA. The watershed extending from 38°34’ N to 39°49’ N and 85°24’ W to 86°31’ W spreads over 6,259 square kilometers, and is a subwatershed in the Ohio River basin,

delineated based on unregulated USGS streamflow station 03371500 (East Fork White River near Bedford, Indiana). The study area predominantly includes forested land followed by agricultural land. Figure 3.1 shows a map of the study area with the delineated stream network. The choice of the study area was motivated by the fact that drought analyses need to be conducted for unimpaired watersheds, where streamflows have not been influenced by upstream diversions, dams, or storage reservoirs.

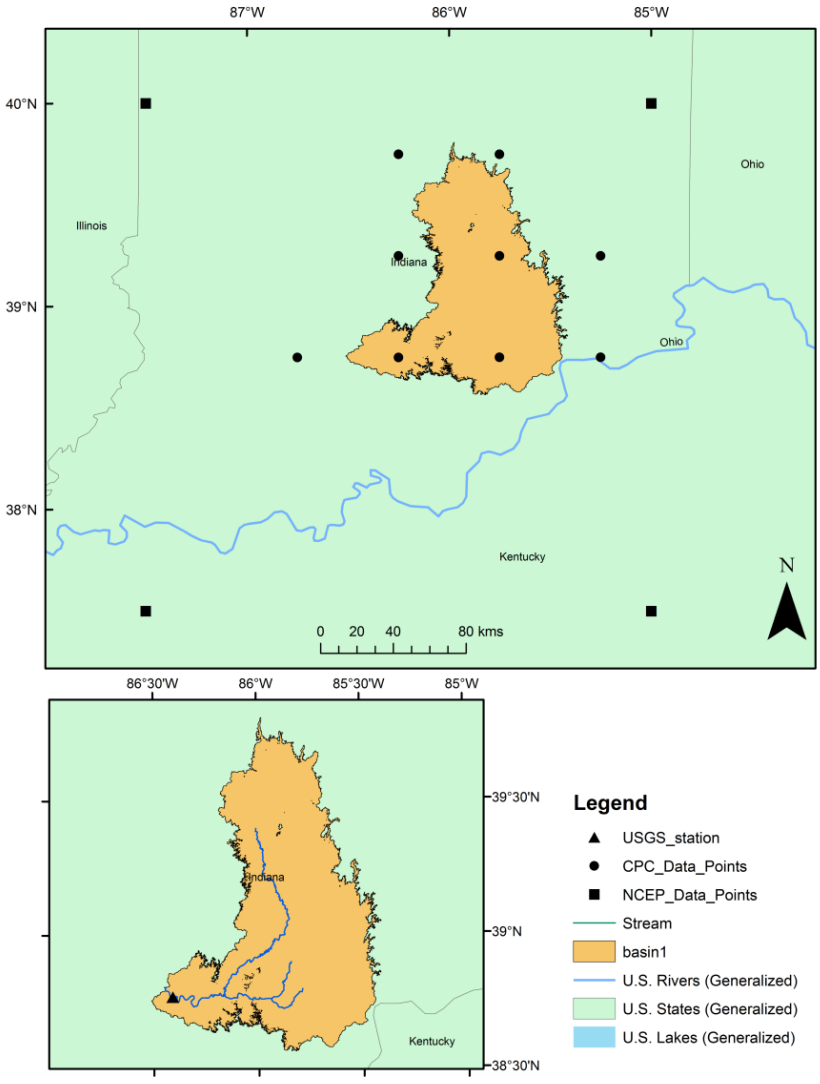

Figure 3.1 Map of study watershed and data points

3.3.2 Data Used

Streamflows depend on many variables. Over the US region, Huang et al. [1996] had identified precipitation, temperature, runoff, and evaporation as variables of interest at a basin scale for soil moisture modeling studies. Along similar lines, soil moisture, precipitation, temperature, runoff, and evaporation were identified as relevant variables for streamflow prediction over the study area. These data sets have been made available by National Weather Service (NWS)-Climate Prediction Center (CPC) established by National Oceanic and Atmospheric Administration (NOAA). Out of these five variables, precipitation and temperature are observational records. Runoff has been calculated from the observed precipitation, using the procedure described in Georgakakos [1986]. Evaporation was obtained using temperature records and by using the relationship in Thornthwaite [1948]. Further, soil moisture was estimated by Huang et al. [1996] using the leaky bucket model. Nearly 10 grid points were identified relevant for the study area at a resolution of 0.5^{o} x 0.5^{o} and monthly CPC data from 1958 to 2010 were used for the first five variables listed in Table 3.1. Monthly streamflow data from 1958 to 2010 recorded at the USGS 03371500 (Figure 3.1) were utilized in this study. Two other variables - mean sea level pressure and wind speed - were also included in the analysis to examine the capability of the graphical modeling technique in identifying and discarding extraneous variables. The National Center for Environmental Prediction-National Centre for Atmospheric Research (NCEP-NCAR) reanalysis 1 project data [Kalnay et al., 1996] were used as proxy data for mean sea level pressure, and zonal (U-wind) and meridional (V-wind) wind speeds near the surface. The variable wind speed data (resultants of zonal and meridional winds) were utilized in the analysis. The monthly data for the variables

used in the study were obtained for the period 1958-2010 at a grid based resolution of 2.5^o x 2.5^o, for four relevant grid points around the study area. The grid locations are shown in Figure 3.1.

Table 3.1 List of variables considered in the analysis

Sl. No	Variables Used	Unit	Period
1	Soil moisture	mm	1958-2010
2	Precipitation	mm	1958-2010
3	Temperature	°C	1958-2010
4	Runoff	mm	1958-2010
5	Evaporation	mm	1958-2010
6	Sea level pressure	mbar	1958-2010
7	Wind Speed	m/s	1958-2010
8	Streamflow	m^3/s	1958-2010

3.4 Methodology

The method requires data processing and quantification of conditional independence structure between different variables using graph theory. Details regarding initial data processing and the adopted graphical modeling technique are provided in this section.

3.4.1 Data Processing

Data for different variables of interest are available at various grid points within and in the neighborhood of the watershed being studied (Figure 3.1). Using thiessen polygon approach, grid station data were averaged over the entire study area to obtain a single monthly time series for each variable. The monthly anomaly time series were

constructed from all variables by subtracting their respective monthly means. With appropriate transformations, it was ensured that the predictors and predictand follow a normal distribution, so as to identify the connections using a Gaussian graphical model, and thus the potential predictors for developing a probabilistic streamflow forecasting model.

3.4.2 Graphical Models

3.4.2.1 Identifying the Conditional Independence Structure

Conditional independence is the cornerstone of graphical modeling technique, offering ease of interpretation and application [Lauritzen, 1996]. In this study we consider use of Gaussian graphical models, i.e. multivariate Gaussian distributions defined on undirected graphs, where the nodes denote variables and the edges provide an idea of statistical dependence structure [Malioutov et al., 2006]. A Gaussian graphical model is therefore an undirected graph $G = (V; E)$ where V is the set of nodes (or vertices) and E is the set of edges connecting pairs of jointly Gaussian variables. Specifically, Gaussian graphical models facilitate the development of sparse and statistically sound models for forecasting applications [Bach and Jordan, 2004]. Several steps are involved in identifying the conditional independence structure for a multivariate Gaussian distribution and are listed below [see Edwards, 2000; Whittaker, 2009].

Let $X = (X_1, X_2, \ldots, X_k)$ be a k-dimensional multivariate Gaussian random variable with mean vector $\vec{\mu} = (\vec{\mu}_1, \vec{\mu}_2, \ldots, \vec{\mu}_k)$ and covariance matrix J such that $J_{ij} = Cov(X_i, X_j)$. In the present context, the vector X includes both the predictand and predictor variables. For a finite data set of size N, the sample mean and variance may be denoted by $\hat{\vec{\mu}}$ and $\hat{\vec{V}}$. Let S denote the inverse of the sample covariance matrix (also called as precision matrix). The precision matrix is rescaled, so that each row is divided by the corresponding diagonal element.

$$S_c(i, j) = S(i, j) / S(i, i) \tag{3.1}$$

The off-diagonal elements of the rescaled precision matrix are set to zero if they are smaller than a specified threshold. The threshold value chosen for pruning the inverse scaled precision matrix was adopted corresponding to a 5% significance level for the length of the record. The information stored in scaled precision matrix is used to construct the conditional independence graph, such that a zero term in the scaled precision matrix corresponds to the absence of an edge between two variables [Dempster, 1972]. The variables that share strong relationships with the predictand variable are shown in the graph. Once the conditional independence graph between different variables is obtained, the next step is to determine the connection strengths. In the case of k-dimensional multivariate Gaussian random variable $X = (X_1, X_2, X_3, \ldots X_k)$, the information divergence for measuring the conditional independence of X_1 and X_2 when $(X_3, \ldots, X_k)$ are given, for instance, is a simple function of the partial correlation between X_1 and X_2 when $(X_3, \ldots, X_k)$ are given. This conditional information is expressed as:

$$Inf(X_1, X_2 \mid X_3, \ldots, X_k) = -0.5 \log\{1 - corr^2(X_1, X_2 \mid X_3, \ldots, X_k)\} \quad (3.2)$$

This expression is based on the concept of Kullback-Leibler information divergence [Cover and Thomas, 1991] between two density functions and Shannon entropy. Equation (3.2) gives a measure of the strength of a connection in the independence graph. For visualization purposes, stronger connections (strength greater than the threshold) between variables are denoted by thick black lines and weaker ones by lighter shaded edges.

3.4.2.2 Model Performance on Synthetic Data

The performance of the proposed graphical model-based predictor selection was first evaluated using a test case whose conditional independence structure is known a priori. In this example, Y is the predictand, and variables X_1, X_2, X_3, X_4 are randomly generated from different Gaussian distributions (mean and standard deviation are given in the parenthesis), and X_5, X_6 are functions of X_3, X_1 respectively. Predictor variables for Y are X_3, X_4 and X_6:

$$\begin{aligned} &\text{Predictors: } X_1 = N(120,15),\ X_2 = N(479,47); \\ &\qquad X_3 = N(30,5),\ X_4 = N(300,134),\ X_5 = {X_3}^2,\ X_6 = \sqrt{X_1} \\ &\text{Predictand: } Y = 14X_3 + 12X_4 - X_6 + 8 \end{aligned} \quad (3.3)$$

The pruned inverse scaled precision matrix for this test case identified three predictors: $X_3, X_4,$ and X_6. Even though X_5 is a function of X_3, and X_1 is a function of X_6, and are considered in the initial predictor set, the graphical model-based predictor selection algorithm discarded X_1 and X_5 in the presence of variables X_6 and X_3 respectively, implying conditional independence of the selected inputs.

For identifying predictors, Sharma [2000] utilized PMI on five synthetic stochastic linear models and two non-linear models. The stopping criterion for the predictor selection algorithm comprised of ascertaining whether the computed PMI was greater than the 95th percentile randomized sample PMI. However, for the synthetic non-linear two variable Threshold Autoregressive order 2 model (TAR 2) in their test data set, the PMI criterion could not correctly identify the predictors, as the method selected an additional predictor, as observed by Sharma [2000]. Since this was the most challenging synthetic data, the TAR 2 model was selected for testing the graphical modeling technique. The TAR 2 model is given by:

$$\begin{aligned} x_t &= -0.5x_{t-6} + 0.5x_{t-10} + 0.1e_t \quad \text{if} \quad x_{t-6} \le 0 \\ &\quad 0.8x_{t-10} + 0.1e_t \qquad\quad \text{if} \quad x_{t-6} > 0 \end{aligned} \tag{3.4}$$

where x_t is a non-linear time series, and e_t is Gaussian noise. The predictor set consists of 15 previous values of x_t (i.e. $x_{t-1}, x_{t-2}, \ldots, x_{t-15}$). Additional details of this synthetic data set can be obtained from Sharma [2000].

The pruned inverse scaled precision matrix for this model identified two predictors: x_{t-10} and x_{t-6} in decreasing order of their connection strengths. The threshold value for pruning the graph was based on the length of the data, as described in the previous section. PMI-based selection had wrongly identified an additional predictor for this test case, and calculation of the 95th percentile randomized sample statistic required substantial computing effort as it involved bootstrapping the predictand variable numerous times (~100) to determine the 95th percentile confidence limits.

Results of graphical model-based predictor selection for the above model indicated that the proposed methodology was better, computationally efficient, and accurate in identifying predictors when compared to PMI-based algorithm.

3.4.3 Streamflow Prediction Modeling

Using the proposed graphical modeling approach, predictors were identified for the monthly streamflow anomaly prediction model. Datasets for calibration period were used to identify the structure of conditional independence graph between all the variables. After pruning the graph by using only variables connecting to streamflow anomaly, the final subset of variables formed the dataset for a parsimonious prediction model. For notational convenience, the predictand variable is labeled as Y, and the remaining variables in the reduced set as $\vec{X}_{sel}$. The conditional independence structure implies Y is independent of $\vec{X} - \vec{X}_{sel} - Y$ given $\vec{X}_{sel}$. Since the ordering of variables is arbitrary, let the reduced predictor set be denoted as:

$$\vec{X}_{sel} = (X_1, X_2, ..., X_r) \tag{3.5}$$

where $r \leq k-1$ and indicates the degree of dimensonality reduction achieved by the conditional graph. The performances of statistical models incorporating the whole predictor set $\vec{X} - Y$, and selected predictors $\vec{X}_{sel}$ were compared to establish the merits of using graphical models as a means of parsimonious selection of input variables.

3.4.4 Statistical Models for Streamflow Prediction

The regression model for streamflow prediction used in the present study is variational relevance vector machines (VRVM). Additionally, ANNs and ARMAX models were used to compare performance of the parsimonious and non-parsimonious models. The performance statistics-coefficient of determination (R^2), Nash-Sutcliffe efficiency (E), and root mean square error (RMSE) were employed to judge the predictive capabilities of the models.

VRVM differs from standard RVM in its complete Bayesian treatment of RVM using principles of variational inference. A brief description of VRVM model is provided here, further details of which can be obtained from Bishop and Tipping [2000] and Tripathi and Govindaraju [2007, 2011]. Given N observations of a set of input vectors $X = \{\mathbf{x}_i\}$ and output $Y = \{y_i\}$ where $i = 1,\ldots,N$ such that $\mathbf{x}_i$ denotes the i^{th} observation in a d-dimensional space, i.e. $\mathbf{x}_i = [x_{i1},\ldots,x_{id}]$, the predictand-predictor relationship in VRVM framework can be represented as

$$y_i = f(\mathbf{x},\mathbf{w}) + \varepsilon = \sum_{m=1}^{M} w_m \phi_m(x_i) + \varepsilon = \mathbf{w}^T \Phi + \varepsilon \tag{3.6}$$

where $\varepsilon \sim N(\varepsilon \mid 0, \sigma_\varepsilon{}^2)$ is the Gaussian error term with mean zero and variance $\sigma_\varepsilon{}^2$ (with precision $\tau_\varepsilon = \sigma_\varepsilon{}^{-2}$). $\mathbf{w} = \{w_m\}$ are the weights associated with the basis functions $\Phi = \{\phi_m\}$, and $m = 1,2,\ldots,M$. The non-linear basis function or the kernel function $\mathbf{K}$, chosen in this application is a Gaussian or radial basis function (RBF) defined as:

$$\phi_{ij} = K(x_i, x_j) = \exp\left(-\sum_{k=1}^{d} \frac{\left(x_{ik} - x_{jk} \right)^2}{{\sigma_{jk}}^2} \right) \tag{3.7}$$

where $\overrightarrow{\sigma_j} = [\sigma_{j1}, \ldots, \sigma_{jd}]$ is the width of the RBF kernel, which is assumed to be constant for all $K(\bullet, \bullet)$, and hereafter referred to as the kernel width parameter σ_{ker} .

The conditional distribution of the output variable given the input vector is Gaussian, and hence, the likelihood of the data set is of the form

$$p(y \mid \mathbf{x}, \mathbf{w}, {\sigma_{\text{ker}}}^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi {\sigma_{\text{ker}}}^2}} \exp\left(-\frac{1}{2{\sigma_{\text{ker}}}^2} \left[y_i - f(\mathbf{x}_i, \mathbf{w}) \right]^2 \right) \tag{3.8}$$

The model bias and weight vectors are then assigned prior distributions (hierarchical priors) of the form

$$p(w_m, \alpha_m) = N(w_m \mid 0, {\alpha_m}^{-1}) \text{ where } m = 1, \ldots, M \tag{3.9}$$

where α_m is the hyperparameter assigned to w_m. Unlike the standard RVM, VRVM assigns hyperpriors for each of the hyperparameters and noise variance:

$$\begin{aligned} p(\alpha_m) &= \text{Gamma}(\alpha_m \mid a_0, b_0) \text{ where } m = 1, \ldots, M \\ p(\tau_\varepsilon) &= \text{Gamma}(\tau_\varepsilon \mid c_0, d_0) \end{aligned} \tag{3.10}$$

The priors were made non-informative to avoid initial bias, by setting $a_0 = b_0 = c_0 = d_0 = 10^{-3}$ [Tripathi and Govindaraju, 2011]. In the above described Gaussian kernel VRVM model framework, the predictive distribution is given by

$$P(y \mid \mathbf{x}, X, Y) = \iint P(y \mid \mathbf{x}, \mathbf{w}, \tau_\varepsilon) P(\mathbf{w}, \tau_\varepsilon \mid X, Y) \, d\mathbf{w} \, d\tau_\varepsilon \tag{3.11}$$

The true posterior $P(\mathbf{w}, \tau_\varepsilon \mid X, Y)$ is then approximated by factorizing the joint distribution of parameters into independent distributions denoted by Q, using variational

principles (i.e., $P(\mathbf{w}, \tau_\varepsilon \mid X, Y) \approx Q(\mathbf{w}, \tau_\varepsilon) = Q_w(\mathbf{w})\, Q_{\tau_\varepsilon}(\tau_\varepsilon)$). Upon further simplifications that are detailed in Bishop and Tipping [2000], the predictive distribution in Eq. (3.11) becomes

$$\begin{aligned} P(y \mid \mathbf{x}, X, Y) &= \int P(y \mid \mathbf{x}, \mathbf{w}, \langle \tau_\varepsilon \rangle)\; Q_w(\mathbf{w})\, d\mathbf{w} \\ &= N(y \mid \mu_w{}^T \phi(x), \sigma^2) \end{aligned} \tag{3.12}$$

$$\text{where } \sigma^2(x) = \frac{1}{\langle \tau_\varepsilon \rangle} + \phi(x)^T \sum_w \phi(x)$$

The prediction was extended to a testing period to validate the calibrated models for the study area. Results from the VRVM-based streamflow prediction model were then utilized in preparing probabilistic forecasts of droughts.

3.5 Results and Discussion

Seven different variables (see Table 3.1) that are likely to influence future streamflows were identified to form the pool of predictors, and the lumped (averaged over the entire watershed) monthly time series for all the variables were prepared for the period 1958 to 2010. Further, streamflow values at the current time step were also considered as an influencing variable for the forecast models at lead times ranging from one to four months. The models were calibrated and tested using data sets from Jan 1958-Dec 1993 and Jan 1994-Dec 2010 respectively. The monthly anomaly series were computed for all the variables by subtracting the corresponding monthly mean values obtained from the calibration period data. The reason for working with anomalies was to develop a model that would do better than simply the long term mean. The lumped monthly time series for each variable anomaly was tested for fit to a Gaussian distribution

using Kolmogorov-Smirnov (KS) test at a 5% significance level, for both calibration and testing periods. In the present study, a two-step approach was adopted for transforming non-normally distributed continuous variables [Templeton, 2011]. Firstly, the variable was transformed into a percentile rank, resulting in uniformly distributed probabilities. The second step applied the inverse-normal transformation to the results of the first step to form a variable consisting of normally distributed z-scores.

3.5.1 Graphical model-based predictor selection

The graphical modeling technique was used to reveal the dependence patterns between anomalies of streamflows and predictor hydroclimatic variables at monthly time step for the calibration period. Four separate graphical models were developed for the four forecasting horizons (1 to 4 months) using the calibration period (1958-1993) data. The threshold value chosen for pruning the inverse scaled precision matrix was adopted as 0.0863 corresponding to a 5% significance level for the length of the record. The graph obtained for the one month-ahead prediction model is shown in Figure 3.2. Predictors relevant for streamflow prediction are outlined by thick boxes, and thick dark connecting lines represent a strong connection between the two variables at the ends.

The graph summarizes the interactions that manifest within transformed anomalies of different variables. Since the main objective was to identify predictors and model streamflows at lead time of one month (predictand), the focus was on the association between streamflow anomaly (SF_{t+1}) and the other predictors. Figure 3.2 reveals that prediction of one-month ahead streamflows is highly influenced by precipitation (PPTN) and soil moisture (SMTR) anomalies with maximum connection

strengths. The connection strengths between streamflow anomaly, and rest of the predictor anomalies are negligible. Figure 3.2 further suggests that, given anomaly values of precipitation and soil moisture, streamflow anomaly is independent of the remaining predictors thereby resulting in a parsimonious model construction.

Table 3.2 lists the selected predictors for the two-, three- and four-months forecast models (graphs not shown for brevity). In case of two-month ahead streamflow forecast, soil moisture, precipitation and runoff anomalies possess strong connections with streamflow anomaly. Thus, parsimony could be achieved even as the streamflow forecasting time horizon changed to two months, but at a reduced level compared to one-month lead time. For a three-month time horizon, significant connection strengths were observed between anomalies of streamflows and soil moisture, runoff, temperature, and evaporation. In the case of four-month ahead forecasts, only soil moisture anomaly shows significant connection with streamflow anomaly (Table 3.2).

Table 3.2 Graphical model-based predictor selection for the four streamflow forecast models

Forecast model→	Selected Predictors $\vec{X}_{sel}$			
	1 month (SF_{t+1})	2 months (SF_{t+2})	3 months (SF_{t+3})	4 months (SF_{t+4})
Streamflow anomaly (SF_t)				
Soil moisture anomaly ($SMTR_t$)	✓	✓	✓	✓
Precipitation anomaly ($PPTN_t$)	✓	✓		
Temperature anomaly ($TEMP_t$)			✓	
Runoff anomaly (RNFt)		✓	✓	
Evaporation anomaly ($EVPN_t$)			✓	
Sea-level pressure anomaly ($PSSR_t$)				
Wind speed anomaly ($WIND_t$)				

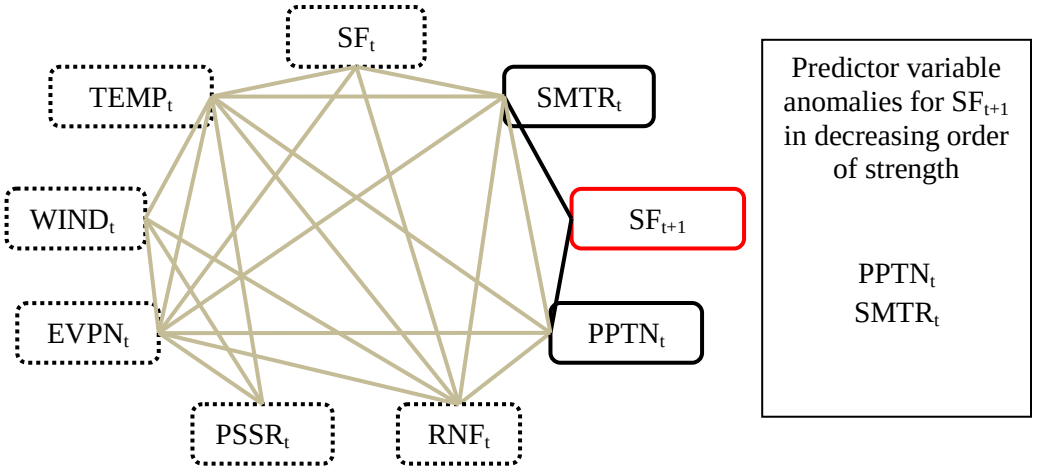

Figure 3.2 Graphical models for one month-ahead monthly streamflow anomaly prediction . Thick black lines and boxes indicate connections and predictors, respectively, relevant for streamflow prediction in the watershed. SF_{t+1} is streamflow anomaly at one-month lead time; SF_t, $PPTN_t$, $SMTR_t$, $TEMP_t$, RNF_t, $EVPN_t$, $PSSR_t$, and $WIND_t$ represent anomalies of streamflows, precipitation, soil moisture, temperature, runoff, evaporation, pressure and wind speed, respectively, at current time step t

While the graphs and connection strengths establish conditional independence relationships and help identify the reduced predictor sets, they do not necessarily reveal the structure of the model to be employed, nor do they indicate the level of performance that will be achieved by these models. However, some preliminary insights are offered by the graph. Figure 3.2, for instance, reveals that the watershed behaves as a reservoir (linear or otherwise) for a one-month time horizon. The output (streamflows) is entirely determined by the input (precipitation), and storage (proxied by soil moisture), and hence hydrologic reservoir models might offer an alternative for modeling streamflows. At two- and three-month lead times, more variables surfaced as necessary inputs offering less dimensionality reduction compared to one-month lead time (Table 3.2). Soil moisture anomaly was found to have a connection for the four-month time horizon, suggesting that of all predictors being considered, soil moisture possesses the longest memory. However,

it is unlikely that a good model for streamflows can be built on soil moisture anomaly alone, and such a model would provide at best only a marginal improvement over the long-term mean. Thus, the utility of the various predictors to update streamflow forecasts beyond the long-term mean decreases with increasing time horizons and establishes our limits of predictability. While the focus was on streamflows, the graphs also reveal the nature of the connections between other input variables. This information would be useful in other hydroclimatic studies.

The variable anomalies that share connections with streamflows were also ranked in decreasing order of their connection strengths and have been shown in Figure 3.2. For one month-ahead streamflow forecasts, precipitation and soil moisture anomalies have the highest rankings with nearly equal strengths. Precipitation anomaly is expected to be a strong predictor as precipitation is the primary driving force for streamflow generation, and the strong role of antecedent moisture conditions is reflected in the equally high rank for soil moisture anomaly. As the lead time increases to two months, graphical models revealed that streamflow anomaly is no longer dominated by precipitation and soil moisture anomalies alone (even though they are ranked among the strongest predictors), as runoff anomaly also comes into play. The other variables: anomalies of temperature and evaporation become significant predictors at longer forecast horizons. However, beyond a forecast horizon of 1 or 2 months, the model prediction capabilities were poor.

The feature selection capability of the graphical modeling approach was compared with the state-of-the-art PMI-based approach, using two stopping criteria: (a) 95^{th} percentile randomized sample statistic [Sharma, 2000], and (b) the Hampel-based score [May et al., 2008; Fernando et al., 2009]. The 95^{th} percentile randomized sample

statistic measure listed in Table 3.3 suggests that the variable is to be selected as a potential predictor when its PMI is greater than this threshold value. The results using PMI on the calibration period data with criterion (a) suggested that the entire set of predictors would be selected in this instance (see Table 3.3), thereby achieving no dimensionality reduction for any of the forecasting horizons. Another disadvantage of the PMI-based method was the computational time involved in the 95^{th} percentile randomized sample statistic estimation.When using PMI along with criterion (b), as the variables are correlated with each other, predictor identification was thwarted by the masking effect that was also noted by previous researchers [May et al., 2008; Fernando et al., 2009]. While PMI-based methods are useful for predictor selection, the present study found the graphical model approach to be more effective for the hydroclimatic data set pertaining to the watershed.

Table 3.3 Details of stepwise predictor selection using PMI criterion

Variable anomaly	Model 1		Model 2		Model 3		Model 4	
	PMI	95^{th}PMI*	PMI	95^{th}PMI*	PMI	95^{th}PMI*	PMI	95^{th}PMI*
SF	0.155	0.016	0.066	0.020	0.036	0.021	0.042	0.020
SMTR	0.166	0.020	0.080	0.022	0.054	0.020	0.053	0.017
RNF	0.151	0.020	0.050	0.022	0.038	0.019	0.048	0.017
TEMP	0.061	0.019	0.062	0.019	0.054	0.020	0.044	0.019
PPTN	0.140	0.018	0.043	0.018	0.028	0.021	0.027	0.019
EVPN	0.079	0.018	0.070	0.018	0.053	0.019	0.035	0.019
PSSR	0.063	0.019	0.065	0.021	0.045	0.019	0.036	0.022
WIND	0.051	0.018	0.045	0.017	0.044	0.018	0.033	0.019

* denotes the 95^{th} percentile randomized sample PMI score

3.5.2 Streamflow Prediction

Once the nature of conditional independence was revealed between anomalies of streamflows and different hydroclimatic variables using the graphical modeling approach, the set of variables with strong connections to streamflow anomaly were considered for the second objective of developing a parsimonious model for predicting streamflows at each of the different lead times. Streamflows were predicted for the four lead time horizons (1 to 4 months) in the following two ways: (i) using all the variables in $\vec{X}-Y$ as predictors, and (ii) using the reduced variable set $\vec{X}_{sel}$, consisting of selected predictors. The model first predicted the streamflow anomaly, which was then converted into streamflows. The coefficient of determination (R^2), Nash-Sutcliffe Efficiency (E), and root mean square error (RMSE) values obtained between predicted and observed values of streamflows for both calibration and testing periods for the four forecasting horizons are listed in Table 3.4. For all the four lead times, the performance evaluation measures calculated for the VRVM model for the two cases- using $\vec{X}-Y$ and $\vec{X}_{sel}$ as predictors are very similar. The R^2, E and RMSE calculated for the parsimonious ($\vec{X}_{sel}$-based) one-month lead time forecast model is 0.69, 0.48 and 81.3 respectively, and those are close to results of the $\vec{X}-Y$-based model (0.71, 0.5, and 79.6, in Table 3.4) during calibration. In certain instances (lead times = 1, 3 and 4 months), it is observed that the parsimonious $\vec{X}_{sel}$-based model outperforms the $\vec{X}-Y$-based model during the testing period.

In addition to using VRVM, popular statistical tools such as ANN and ARMAX were used to compare the performance of $\vec{X}-Y$ and $\vec{X}_{sel}$-based models. ANN regression model adopted in the study comprised of input nodes corresponding to $\vec{X}-Y$ and $\vec{X}_{sel}$, streamflow anomaly as the output node, and n hidden neurons that are arranged in the hidden layer. The neurons in different layers interact with each other via weighted connections. A feed-forward network ANN was used in the present study, using Levenberg-Marquardt backpropagation scheme as the learning algorithm. The third model: ARMAX (p,q,r) consists of p autoregressive, q moving average, and r exogenous input predictor terms. While the autoregression model (AR) specifies the dependence of the output variable Y on its value at previous p time steps, the moving average (MA) part is a linear regression of current and previous q white noise error terms. The white noise error terms are normally distributed. In this study, the predictors in $\vec{X}-Y$ and $\vec{X}_{sel}$ are the exogenous inputs in the ARMAX model. The R^2, E and RMSE values calculated for these two regression techniques for all the forecast models are provided in Table 3.4. During ANN model calibration, while the $\vec{X}-Y$-based model results are slighly superior to $\vec{X}_{sel}$-based models, they performed equally well during the testing period. Whereas using ARMAX regression, especially for the 1-, 2-, and 4-months forecasts, the parsimonious models performed as well as the $\vec{X}-Y$-based model. At two, three and four-months lead times, models with autoregressive lags of 1, 4 and 1, respectively, performed better. However, the best performing parsimonious ARMAX models at all lead times agree with the graphical model-based predictor selection; they do

not use any previous month streamflows. In Table 3.4, also provided are the RBF kernel width σ_{ker}, the number of hidden neurons in the hidden layer n, and the AR and MA lags $\{p,q\}$ corresponding to the best VRVM, ANN and ARMAX models, respectively. The selection criterion for σ_{ker} for VRVM was based on achieving high variational lower bound value while preserving good generalization capabilities [Tripathi and Govindaraju, 2011]. The results from different regression techniques for $\vec{X}-Y$- and $\vec{X}_{sel}$-based models indicate that given the reduced set of predictor variables with strong connections $\vec{X}_{sel}$ through conditional independence, no extra information from other variables $\vec{X}-\vec{X}_{sel}-Y$ was needed to improve streamflow prediction performance using the three models, thus resulting in parsimonious models.

The prediction of streamflow anomaly using the hydroclimatic precursor anomalies selected by the graphical model is expected to be more reliable if the selected variables have relatively high connection strengths. In the present study, models for one-month lead time forecasts revealed strong dependence patterns between anomalies of streamflows and selected hydroclimatic variables. As expected, results in Table 3.4 indicate that relatively more confidence can be placed in making streamflow predictions for one-month lead time when compared to other longer lead times. To explore other applications, further study was restricted to streamflow predictions for only one-month lead time. In this case, two hydroclimatic variables (precipitation and soil moisture) are identified by the graphical modeling approach as exhibiting strong connections with streamflow anomaly (Figure 3.2).

Table 3.4 Coefficient of determination (R^2), Nash-Sutcliffe efficiency (E) and root mean square error (RMSE, in cumecs) values for comparing calibration and validation performance of monthly streamflow prediction models: VRVM, ANN and ARMAX using all hydroclimatic predictors, and using parsimonious models (GM-VRVM, GM-ANN, and GM-ARMAX) at lead times - 1 to 4 months

			VRVM (σ_{ker}=200)	GM-VRVM (σ_{ker}=200)	ANN (n=3)	GM-ANN (n=3)	ARMAX (p=0,q=6)	GM-ARMAX (p=0,q=2)
lead time = 1 month	Calibration	R^2	0.71	0.69	0.74	0.70	0.70	0.70
		E	0.50	0.48	0.55	0.49	0.49	0.49
		RMSE	79.60	81.30	75.40	79.90	79.96	80.04
	Validation	R^2	0.61	0.62	0.58	0.61	0.62	0.62
		E	0.33	0.36	0.32	0.36	0.37	0.36
		RMSE	117.80	115.70	119.27	115.78	114.8	115.27
			VRVM (σ_{ker}=220)	GM-VRVM (σ_{ker}=150)	ANN (n=4)	GM-ANN (n=3)	ARMAX (p=1,q=1)	GM-ARMAX (p=0,q=1)
lead time = 2 months	Calibration	R^2	0.63	0.62	0.66	0.63	0.63	0.63
		E	0.40	0.38	0.43	0.40	0.40	0.40
		RMSE	87.20	88.60	84.76	87.48	86.61	86.79
	Validation	R^2	0.47	0.46	0.46	0.46	0.48	0.48
		E	0.16	0.16	0.15	0.15	0.17	0.18
		RMSE	131.90	132.10	132.92	133.15	130.92	130.85
			VRVM (σ_{ker}=220)	GM-VRVM (σ_{ker}=150)	ANN (n=2)	GM-ANN (n=2)	ARMAX (p=4,q=3)	GM-ARMAX (p=0,q=1)
lead time = 3 months	Calibration	R^2	0.61	0.61	0.65	0.64	0.64	0.62
		E	0.37	0.37	0.42	0.40	0.40	0.38
		RMSE	88.89	89.50	85.79	86.75	86.48	87.81
	Validation	R^2	0.49	0.50	0.50	0.49	0.50	0.49
		E	0.18	0.19	0.18	0.18	0.19	0.18
		RMSE	131.80	131.09	131.28	131.36	130.55	131.82
			VRVM (σ_{ker}=200)	GM-VRVM (σ_{ker}=255)	ANN (n=2)	GM-ANN (n=2)	ARMAX (p=1,q=1)	GM-ARMAX (p=0,q=3)
lead time = 4 months	Calibration	R2	0.62	0.61	0.64	0.60	0.62	0.62
		E	0.39	0.37	0.40	0.37	0.39	0.38
		RMSE	88.40	89.30	87.05	89.80	87.82	88.37
	Validation	R2	0.49	0.50	0.50	0.49	0.51	0.51
		E	0.17	0.19	0.19	0.18	0.19	0.19
		RMSE	136.15	135.16	135.19	135.27	134.94	134.97

Note: σ_{ker} is the kernel width parameter used in VRVM, n is the number of hidden neurons in ANN model, and p and q are respectively the number of auto-regressive and moving average lags in ARMAX model.

The VRVM streamflow prediction model was used for further analysis because it yields predictive distributions of forecasted streamflows instead of point estimates allowing for probabilistic classification. The calibrated VRVM model for one-month lead time with parsimonious inputs had kernel width σ_{ker} =200, and its performance was evaluated for both calibration (1958-1993) and testing (1994-2010). The R^2 values were 0.69 and 0.62, and RMSE values were 81.3 and 115.7 respectively during calibration and testing periods (Table 3.4). Comparisons between observed and predicted one month-ahead streamflows for some years in calibration and testing (forecasted using the parsimonious model) are shown in Figure 3.3.

The plots indicate that the developed model could capture the trends in flows both during the calibration and testing periods. These plots also show error band of one-standard deviation about the predicted values for the selected years. There is good agreement between observed and predicted monthly streamflows, especially during low flow months in 1981-82, 1984, 1986, 1988, and 1991 as shown in Figure 3.3a. The high flows/peaks in some months (in 1986, 1988 and 1992) matched well with the predicted values. As seen in Figure 3.3b, low flows were predicted well in 2002, and during the years 2005-2009 of the testing period. The predicted flows for the years 1999-2001 closely followed the observed values. There are some observed high flows that are outside the prediction band during the testing period. Flow peaks of such magnitudes occurred only in this time window (Figure 3.3b), and were not present during calibration period causing these discrepancies. Figure 3.3 also contains the plots of inputs to the parsimonious prediction model−monthly precipitation and soil moisture values. As expected, the flow peaks in the years shown are associated with increased monthly

precipitation and soil moisture values. Extremely low flows during the calibration period-for instance, in 1984, 1988, 1992, and during the testing period–in 1999, 2007 and 2010, are well correlated with the low values of both the inputs.

Additionally, the resulting graphical model imparts useful information about the underlying hydrologic model. If we consider only streamflows as the output of interest, low streamflow values are generally associated with baseflow conditions where soil moisture plays a dominant role in determining the fluxes that maintain streamflows. Any precipitation likely pushes the existing soil moisture towards the stream. For peak streamflows, even though these two variables are still the prominent predictors for one-month lead time (Figure 3.3), it is likely that all the non-linearities are not well captured in the prediction model. The implication is that the model is more capable of predicting low streamflow values and is therefore more suitable for conducting drought-related studies. Such a model could serve as a trigger for one-month ahead hydrologic droughts, and would be useful for allocating surface water rights for irrigation purposes.

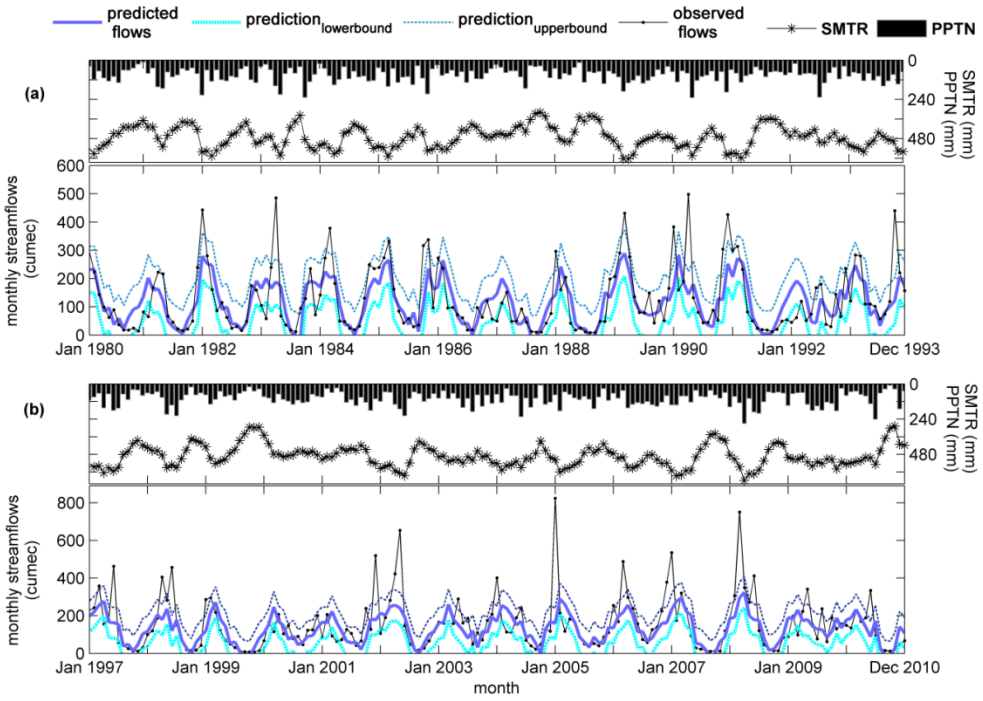

Figure 3.3 Comparison of observed and predicted monthly streamflows during (a) 1980-1993 in the calibration period and (b) 1997-2010 in the testing period. The upper and lower prediction bounds correspond to one standard error of prediction. Inputs to the parsimonious prediction model – monthly soil moisture (SMTR) and precipitation (PPTN) are shown above the respective streamflow plots

3.5.3 Application to Hydrological Droughts

The standardized streamflow drought index was used for drought analysis in the watershed [Shukla and Wood, 2008]. This index is similar to the standardized precipitation index (SPI) calculated for categorizing meteorological droughts [McKee et al., 1993]. A positive value of this index quantifies the degree of wetness, while a negative value indicates the degree of dryness. Table 3.5 presents the drought severity classification suggested by the United States Drought Monitor (USDM; http://drought monitor.unl.edu/classify.htm) for different hydrological conditions ranging from exceptional drought (D4) to normal conditions to exceptionally wet (W4).

Table 3.5 Drought categories and corresponding standardized streamflow drought index range

Drought Category	Description	Range*
D4	Exceptional drought	(-∞ to -2.0]
D3	Extreme drought	(-2.0 to -1.6]
D2	Severe drought	(-1.6 to -1.3]
D1	Moderate drought	(-1.3 to -0.8]
D0	Abnormally dry	(-0.8 to -0.5]
Normal	Normal condition	(-0.5 to 0.5)
W0	Abnormally wet	[0.5 to 0.8)
W1	Moderately wet	[0.8 to 1.3)
W2	Severely wet	[1.3 to 1.6)
W3	Extremely wet	[1.6 to 2.0)
W4	Exceptionally wet	[2.0 to ∞)

*( ) – open ended boundary; [ ] – closed ended boundary

In order to assess the drought forecasting ability of the model, the Heidke skill score (HSS) was computed [Doswell et al., 1990; Jolliffe and Stephenson, 2003; Wilks, 2006]. The HSS gages the accuracy of the model forecast relative to the accuracy of random chance. The range of HSS is -∞ to 1. A score of 0 reflects no skill, a score of 1 is attained with perfect forecasts, whereas, negative scores indicate that chance forecasts are better than the predictions. Table 3.6 provides a quantitative assessment of the drought prediction ability of the model during both calibration and testing periods. During the model calibration, out of a total of 179 droughts observed, the model identified 114 instances. For the testing period, the model predicted drought 32 times out of the 54 observed droughts. The HSS scores shown in Table 3.6 indicate acceptable performance [Barnston, 1992].

Table 3.6 Contingency table showing drought prediction performance during calibration and testing periods

Drought Forecast	calibration*		testing#	
	Drought observed		Drought observed	
	Yes	No	Yes	No
Yes	114	60	32	26
No	65	193	22	124

Note: * Heidke Skill Score: $HSS_{calibration} = 0.41$, # $HSS_{testing} = 0.40$

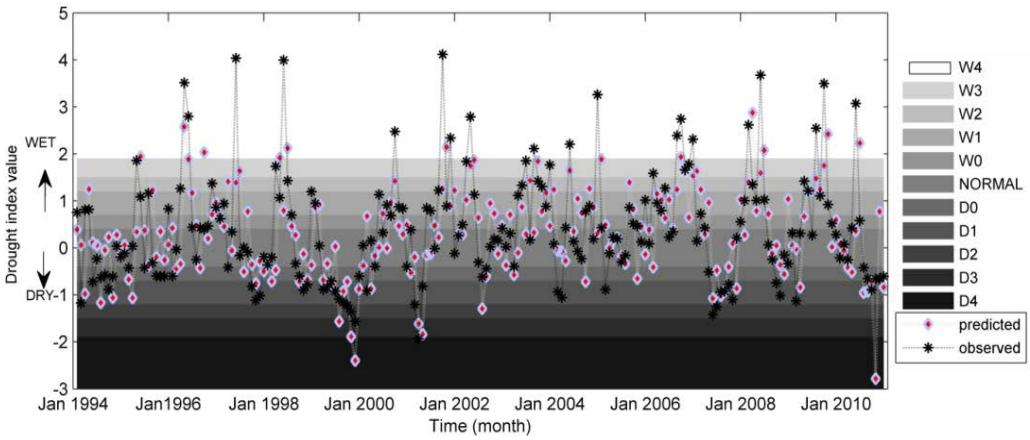

Figure 3.4 Observed and predicted values of standardized streamflow drought index for the model testing period (1994-2010)

The standardized streamflow index values computed from predicted and observed one month-ahead streamflows during the testing period are shown in Figure 3.4. The different drought and wet categories are shown by horizontal shaded bands for easier interpretation. Inspection of Figure 3.4 shows good predictions by the model for the dry periods. The most severe hydrologic droughts were observed in 1994, 1999, 2001, 2007 and 2010, and are predicted by the model too. Continuously dry months predicted during 1999 and the summer of 2001 and 2007 match well with observations. In the testing period, overall, normal-to-wet conditions dominated streamflows; while there were some observed streamflow deficits, the model predicted more droughts, both in number and in severity.

The conditional independence-based model developed in the study was used for analyzing low flow predictions during the testing period. To disaggregate the soil moisture and precipitation anomaly data corresponding to drought and non-drought conditions, the means of these predictors for both the categories were determined. For any new pair of soil moisture and precipitation anomaly data, the Euclidean distances to the centers of drought and non-drought cases were computed as a and b, respectively, as shown in Figure 3.5. During the testing period, streamflows were predicted for all the data sets falling in the drought category (i.e. $a < b$ in Figure 3.5) using the conditional model developed for low values of streamflows during the calibration period. The predicted streamflows were categorized into different drought states based on thresholds in Table 3.5 for probabilistic analysis.

Probabilistic prediction of different drought categories performed during the testing period is shown in Figure 3.6. The height of the designated color bar for each drought category in any drought event reflects the probability of that particular class, thus providing a probabilistic classification and expressing model uncertainty in assigning a drought class to predicted streamflow events. This drought classification was performed only when the precursors—precipitation and soil moisture anomalies from the previous month—suggested drought conditions. Overall, the results are consistent with low streamflow values corresponding to higher probabilities of drought categories.

Whenever a drought was observed, the associated probabilities of drought classes were markedly high compared to non-drought classes. If we examine the exceptional drought (D4) events during the testing period (Figure 3.6), in August and November months of 1999, D4 droughts were predicted by the model with a probability of 44 and

38 percent respectively, during summer of 2001, on an average 30% probability for D4 drought was obtained, and in November 2010, a 43% D4 drought was predicted. In most of these cases, the observed droughts were of similar severity. A smaller number of drought occurrences were reported during the testing period in this watershed. The normal cases were accurately predicted by the model, and a few wet scenarios had a high chance despite existing drought conditions. These differences highlight some of the limitations of the approach, but are also reflective of the level of uncertainty and limits of predictability that can be achieved from one-month hydrologic drought trigger information for this watershed.

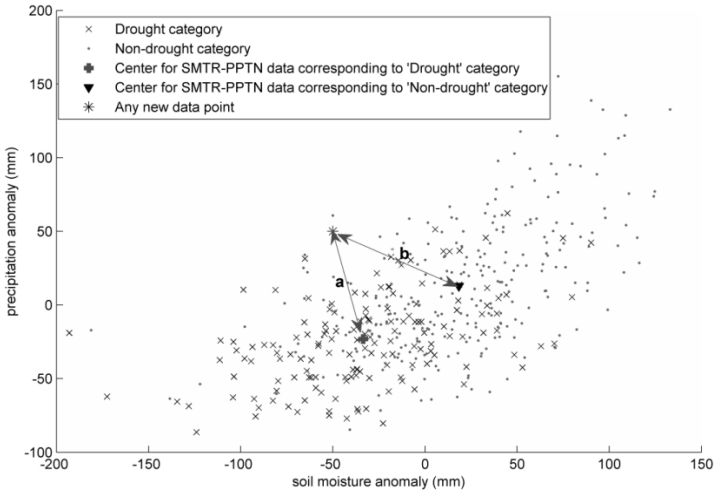

Figure 3.5 Scatter plot between Soil Moisture (SMTR) and Precipitation (PPTN) anomaly data showing centers of ‘Drought’ and ‘Non-Drought’ categories. Whenever a<b (i.e. drought category), a probabilistic prediction of drought categories are made

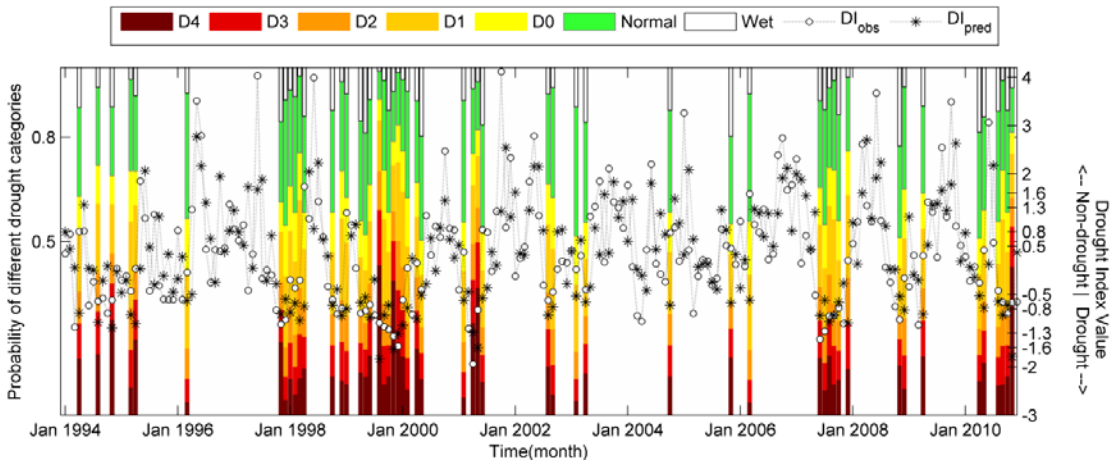

Figure 3.6 Probabilistic prediction of different drought categories during the testing period (1994-2010)

3.6 Summary and Conclusions

This chapter utilized a graphical modeling technique employing conditional independence to address predictor selection in hydroclimatic analysis. Tests with linear and non-linear synthetic data sets demonstrated the dimensionality reduction achieved by this approach. Comparisons with other state-of-the-art predictor selection method based on PMI showed that the proposed graphical modeling approach was more robust and incurred smaller computational burden.

Using the streamflow data for an Indiana watershed, results were examined for four different time horizons using a set of eight variables that are expected to influence the monthly streamflows. The graphs revealed that given precipitation and soil moisture, other variables are not needed for one month-ahead streamflow forecasts, while runoff would be needed for two-month lead time forecasts. Soil moisture and precipitation, apart from being conditionally important as predictors, also possessed the strongest connection

strengths. Temperature, runoff, and evaporation shared an on-and-off relationship with streamflows at longer lead times. However, the long-term mean of streamflows would likely not be improved upon with the help of other hydroclimatic variables for lead times greater than two months.

The graphical modeling approach allowed for development of a parsimonious VRVM-based probabilistic model for prediction of streamflows. The forecast model was used for prediction of streamflows and hydrological droughts over the study area. The prediction performance was evaluated in terms of R^2, E and RMSE values (Table 3.4) and the resulting parsimonious models demonstrated similar performance as the higher dimensional model. The predictive capabilities were equally good while using the parsimonious model during model testing. Results from popular statistical techniques such as ANN and ARMAX yielded similar results. Drought analysis results using a contingency table showed that more than 50% of drought incidents during the calibration and testing periods were successfully captured, indicating overall model robustness. On the other hand, the PMI-based predictor selection had suggested to retain the entire predictor set as shown in Table 3.3.

The following conclusions are drawn from the graphical model-based predictor selection study:

i. The graphical modeling approach utilized here was successful in establishing conditional independence that led to reduced model complexity especially for one-month lead time.

ii. The method allowed development of parsimonious models that were used for conducting an exploratory analysis into droughts.

iii. The general results and conclusions about the importance of soil moisture and precipitation for short-term streamflow predictions are likely to hold for other similar-sized watersheds as well. The same is true for the level of prediction capability at one-month lead times.

iv. However, the specific graphs are likely to be different for different watersheds even for the same time lag, as relative importance of variables for streamflow prediction would depend very much on precipitation rates, travel times and storage capacities of individual watersheds. These properties are known to be scale-dependent, and the evolution of these graphs with spatial scale would allow us to determine how the roles of predictor variables change with scale−a topic of future study.

Overall, this method holds promise for applications in statistical models where predictor selection is of concern, for example, in downscaling studies. The method would serve as a useful first step before construction of complex models is undertaken, especially when physics-based models are either not available or are too complex for practical use. The conditional independence structure would provide useful insights into the construction of models for various hydrologic applications.

CHAPTER 4. PROBABILISTIC ASSESSMENT OF AGRICULTURAL DROUGHTS USING GRAPHICAL MODELS

4.1 Abstract

Agricultural droughts are often characterized by soil moisture in the root zone of the soil, but crop needs are rarely factored into the analysis. Since water needs vary with crops, agricultural drought incidences in a region can be characterized better if crop responses to soil water deficits are also accounted for in the drought index. This chapter investigates agricultural droughts driven by plant stress due to soil moisture deficits using crop stress functions available in the literature. Crop water stress is assumed to begin at the soil moisture level corresponding to incipient stomatal closure, and reaches its maximum at the crop's wilting point. Using available location-specific crop acreage data, a weighted crop water stress function is computed. A new probabilistic agricultural drought index is then developed within a hidden Markov model (HMM) framework that provides model uncertainty in drought classification and accounts for time dependence between drought states. The proposed index allows probabilistic classification of the drought states and takes due cognizance of the stress experienced by the crop due to soil moisture deficit. The capabilities of HMM model formulations for assessing agricultural droughts are compared to those of current drought indices such as standardized precipitation evapotranspiration index (SPEI) and self-calibrating Palmer drought

severity index (SC-PDSI). The HMM model identified critical drought events and several drought occurrences that are not detected by either SPEI or SC-PDSI, and shows promise as a tool for agricultural drought studies.

4.2 Introduction

The onset of an agricultural drought event is typically marked by a decline in the soil moisture level below a threshold value that affects crops. Precipitation, soil moisture, and temperature are the common variables adopted for agricultural drought studies [Mishra and Singh, 2010]. Various indices for characterizing agricultural droughts are listed in Maity et al. [2013]. Among these, Palmer drought severity index [PDSI; Palmer, 1965], crop moisture index [CMI; Palmer, 1968], soil moisture anomaly index [Bergman et al., 1988], and vegetation condition index [VCI; Liu and Kogan, 1996] are popular.

Researchers typically regard soil moisture as the most appropriate indicator of agricultural droughts [Keyantash and Dracup, 2002; Karamouz et al., 2004; Sheffield and Wood, 2008]. Estimation of soil moisture from ground measurements is difficult due to heterogeneity caused by the spatially varying precipitation, land cover, soil and topography [Margulis et al., 2002; Vereecken et al., 2008]. Temporal and spatial resolution of soil moisture is also crucial for predicting adequate soil profile wetting and drying between precipitation events. The role of soil moisture in recurring droughts in North America was studied by Oglesby and Erickson [1989]. Sheffield et al. [2004] used soil moisture estimates from the variable infiltration capacity (VIC) model to develop a drought index that showed major drought events of the past and had good correlations

with PDSI. Lakshmi et al. [2004] found that the deep layer soil moisture was capable of characterizing droughts in the Mississippi River Basin. The soil moisture deficit index (SMDI) developed by Narasimhan and Srinivasan [2005], based on weekly soil moisture deficits, had good correlation with indices such as SPI and PDSI, and offered better performance because of its fine spatial and temporal resolution. The authors used soil water assessment tool (SWAT) model to simulate daily soil moisture values at 4 km X 4 km spatial resolution that were then aggregated to a weekly time scale. Tang and Piechota [2009] explored the possibility of deep layer soil moisture as an indicator of climate extremes, and linked it to PDSI, precipitation, and streamflows. Their study utilized soil moisture as a drought indicator for characterizing the hydrologic status for the Colorado River Basin, and further identified the spatial and temporal variability of soil moisture in response to drought events in the region.

Root-zone soil moisture availability is used by agencies such as the United States Department of Agriculture (USDA)-International Production Assessment Division (IPAD)−as a major factor influencing crop yield forecasts [Bolten et al., 2010]. When Wu et al. [2011] performed drought vulnerability assessment for China, seasonal crop water deficiency, available soil water-holding capacity and irrigation were adopted as the important drought indicators. The soil water holding capacity is a function of soil type, and varies spatially across a region creating patterns of crop water stress and water resource availability. Maity et al. [2013] characterized drought proneness of Malaprabha Basin, India, via a copula model for resilience and vulnerability values calculated from modeled soil moisture data for the region.

Since water needs vary with crops, agricultural drought incidences in a region can be assessed better if crop responses to soil water deficits are also accounted for in the index. Water stress influences rate of photosynthesis and stomatal closure and affects crop production [Scholes and Walker, 1993]. Denmead and Shaw [1960] studied the effect of soil moisture deficit on the development and yield of corn, by imposing soil moisture deficit at different growth stages. Holt et al. [1964] investigated the effect of stored soil moisture at planting on corn yields, and developed regression equations for relating soil moisture to corn yield.

A quantitative understanding of the plant response to water stress requires detailed study of soil moisture dynamics that include soil-water-air interaction, nutrient uptake by plants, and transpiration. Soil moisture deficits directly control the plant water potential that determines transpiration losses and the turgor pressure in plant cells [Porporato et al., 2001]. The role of water stress in the structure and functioning of vegetation in African savannas (grassland ecosystems) was studied by Rodriguez-Iturbe et al. [1999a,b]. The authors proposed a measure of “static” vegetation stress that can be calculated from soil moisture levels corresponding to plant wilting and full turgor. The “static” stress is zero when soil moisture is above the level of incipient stomatal closure (full turgor) and reaches a maximum value of one when soil moisture is at the wilting point of a plant. These two stages are based on the effects of water stress on plant physiology [Hsiao, 1973]. Porporato et al. [2001] later introduced “dynamic” water stress to address the mean intensity, duration and frequency of soil moisture deficits. Laio et al. [2001] developed a stochastic model for soil moisture and water balance studies.

Drought conditions for crops in the Midwest are, by and large, determined by the soil water availability rather than by precipitation or evaporation. The plant response to water stress in the root zone of a soil could be used to develop a new agricultural drought index. Such an index would take due cognizance of crop needs. However, the changing soil moisture status and different crop rotation patterns followed in agricultural fields require that the drought analysis be performed in a statistical sense. A probabilistic assessment would convey the uncertainty in agricultural drought classification that popular indices (SPEI, PDSI, SPI) do not provide. Madadgar and Moradkhani [2013, 2014] developed a probabilistic forecast model for future hydrologic droughts in a Bayesian framework that allows probabilistic predictions and accounts for uncertainty in drought characterization. In this study, agricultural drought events in the state of Indiana are investigated in a probabilistic framework using graphical models−specifically hidden Markov models (HMMs)−given the temporal dependence that exists between drought states. The crop stress function values derived from soil moisture data are used to define agricultural drought states (1-near normal, 2-moderate drought, 3-severe drought, and 4-extreme drought).

Hidden Markov models have been used for solving numerous practical problems in speech processing [Leggetter and Woodland, 1995], signal processing [Crouse et al., 1998], genomics [Yau et al., 2011], tunneling design [Leu and Adi, 2011], meteorological studies [Hocaoğlu et al., 2010] and air quality modeling [Zhang et al. 2012]. Mallya et al. [2013a] utilized HMMs to model meteorologic and hydrologic droughts. Many of these applications used Gaussian emission distributions [Leggetter and Woodland, 1995; Burget et al., 2010; Mallya et al., 2013a]. Alternatively, atmospheric ozone levels were

modeled using Gamma hidden Markov models by Zhang et al. [2012], and Sun et al. [2013] used HMMs with log-normal, Gamma and generalized extreme value (GEV) distributions to predict particulate matter concentrations.

Unlike previous studies [Mallya et al., 2013a; Zhang et al., 2012], the crop water stress function used in this study is bounded between [0,1], and as a result, previously utilized emission distributions are not suitable. This chapter describes a new class of HMMs with beta emission probability distributions. These new models were used for developing probabilistic classification models for agricultural droughts in Indiana. The merits of HMM-based probabilistic agricultural drought index over SPI, self-calibrating PDSI and SPEI were investigated. The organization of rest of the chapter is as follows: section 4.3 describes the study area and data used, section 4.4 explains the methodology adopted in the development of the probabilistic index, followed by results and discussion in section 4.5, and finally the conclusions derived from the study are presented in Section 4.6. In addition, Appendix A provides derivations of equations used in the methodology.

4.3 Study Area and Data Used

4.3.1 Study Area

To examine the applicability of the graphical model, the state of Indiana, USA is chosen as the study area. Indiana is nationally ranked for agricultural production, major cultivated crops being corn and soybean. For instance, Figure 4.1 illustrates the cultivation pattern followed in a small patch of land in Lake County in northern Indiana during the period 2000-2012, where corn and soybean are predominant. Crop rotation,

fallow land, and double cropping practices have been adopted in this area. Winter wheat, alfalfa and pasture grass were grown as minor crops in alternate years. Livestock and dairy farming thrive on agriculture over such farmlands in Indiana and other Midwest states.

Unfortunately, droughts are common in the Midwest, and hamper the prospects of large yields from these farms. Consequences of the recent 2012 drought in US can be found in Mallya et al. [2013b] and Kerr [2012]. Figure 4.2 shows the extent of drought extremes over Indiana evaluated by United States Drought Monitor (USDM) for July 24, 2012. The USDM map identifies regions experiencing different drought categories ranging from D0 (abnormally dry) to D4 (exceptionally dry) for that particular day, and the classification criteria are described in http://droughtmonitor.unl.edu/AboutUs /ClassificationScheme.aspx. More than half of the state was affected by an extreme drought (Figure 4.2). The major impact of agricultural droughts is on crop cultivation in the affected regions. From an economic point of view, droughts have a detrimental effect on corn and soybean prices in Indiana under the current agricultural conditions, and are particularly devastating to livestock producers (http://www.ibrc.indiana.edu/ ibr/2012/outlook/articles/agriculture.pdf).

4.3.2 Data Used

The yearly cropping pattern of Indiana was obtained from Cropland Data Layer (CDL) that is hosted on CropScape [Han et al., 2012; http://nassgeodata. gmu.edu/CropScape/]. The CDL is a raster, geo-referenced, crop-specific land cover data layer created annually for the continental United States using moderate resolution

satellite imagery and extensive agricultural ground truth. It is developed by the National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA). This data is available from 2000-2012 for Indiana. Average crop distribution in acreage for this time window was extended to cover the 1948-2012 period.

For soil moisture data, the Climate Prediction Center's (CPC) 0.5° x 0.5° resolution global monthly datasets [Fan and van den Dool, 2004] were used. The data sets have sufficiently long record lengths needed for robust modeling. Huang et al. [1996] outline the procedure for constructing this monthly soil moisture time series data sets over the entire continental U.S. with a 1600 mm deep one-layer soil moisture model. Their model is based on the water budget in the soil and uses monthly temperature and monthly precipitation as inputs. Estimated evapotranspiration, runoff and groundwater loss used in the CPC soil moisture model are derived from these two inputs. A total of 52 CPC grid points fall over Indiana, and soil moisture data from the period 1948-2012 were extracted at these grid points for this study.

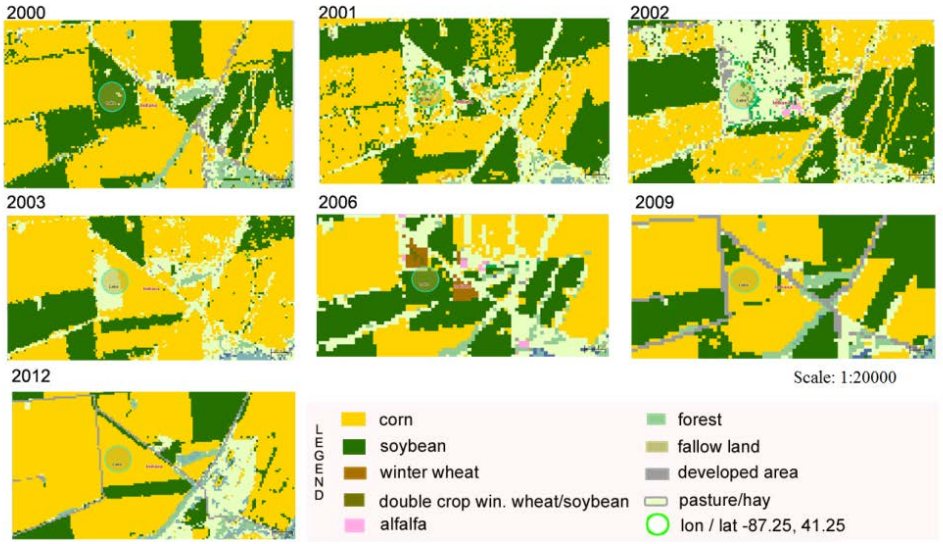

Figure 4.1 Cropping pattern in a small patch of agricultural field in Lake County, Indiana, US during 2000-2012 where the yearly changes in land use and land cover are evident (adapted from http://nassgeodata.gmu.edu/CropScape/)

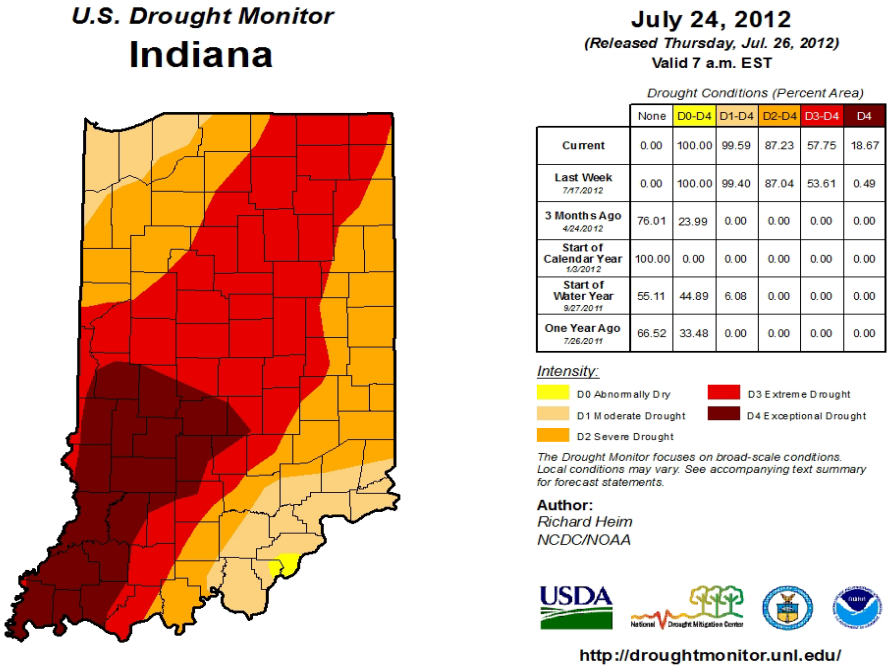

Figure 4.2 Extent and magnitude of 2012 drought in Indiana- in July 2012, one of the hottest months of the year, captured by the U S Drought Monitor with D0 being the least, and D4 being the most intense drought categories listed. (The U.S. Drought Monitor is jointly produced by the National Drought Mitigation Center at the University of Nebraska-Lincoln, the United States Department of Agriculture, and the National Oceanic and Atmospheric Administration. Map courtesy of NDMC-UNL)

4.4 Methodology

Development of an HMM-based probabilistic drought index required estimation of crop water stress, studying the temporal dependence between drought states, choice of emission distribution, parameter estimation, and model selection. These various steps are briefly described in this section.

4.4.1 Estimation of Crop Moisture Stress Function

Plant water potential is controlled by the soil moisture present in the root zone. With excess moisture, the plant water potential increases and turgor in leaves is very high, as a result of which stomatal pores open and evapotranspiration is in full swing. However, under conditions of soil moisture deficit, there is a drop in the water potential in plants, inhibiting their ability to take up water from the soil, as a result of which the stomatal openings close to avoid loss of available water. The sequence of events that take place in plants in response to water stress can be understood based on the varying levels of stomatal closure. Incipient stomatal closure is among the first symptoms, and finally as the plant starts wilting, complete closure would take place. Rodriguez-Iturbe et al. [1999a,b] quantified the plant water stress as a function of soil moisture level in the soil at that instant ("static" water stress) ζ –such that it is zero when soil moisture is above the level of incipient stomatal closure and has its maximum value of 1 when the soil moisture causes wilting (denoted as $s*$ and s_w respectively). Between $s*$ and s_w, the authors suggested a non-linear increase of plant water stress with soil moisture deficit as

$$\zeta(t) = \begin{cases} 1 & \text{for } s < s_w \\ \left[\dfrac{s^* - s(t)}{s^* - s_w} \right]^m & \text{for } s_w \le s(t) \le s^* \\ 0 & \text{for } s > s^* \end{cases} \tag{4.1}$$

where $s(t)$ is the soil moisture content at time t, and m is a measure of the non-linearity desired in the crop water stress model. A value of $m = 2$ is used for crops in this study. The values of s_w, s^* and m vary with plant species.

The crop distribution information at various CPC grid points over the Indiana region were extracted to develop corresponding weights for dominant agricultural crops and multiplied to the crop water stress value of each crop. The resulting monthly effective crop stress time series at each grid point was used for agricultural drought analysis in the region.

The crop distribution information at various CPC grid points over the Indiana region were extracted to develop corresponding weights for dominant agricultural crops and multiplied to the crop water stress value of each crop. The resulting monthly effective crop stress time series at each grid point was used for agricultural drought analysis in the region.A hidden Markov model (HMM) was used to develop a probabilistic classification model to define agricultural droughts. A schematic of the graphical model used in the study is shown in Figure 4.3. It illustrates the concept of estimating crop stress ζ using soil moisture and crop information. The non-linear increase in ζ between s^* and s_w is represented in the graph in Figure 4.3. The HMM graph structure with the hidden drought states (in dashed boxes) is shown in the same figure. In this approach, a certain range of crop water stress values define a drought state,

and the range varies spatially. Brief description of the theory of HMMs is provided in subsequent sections.

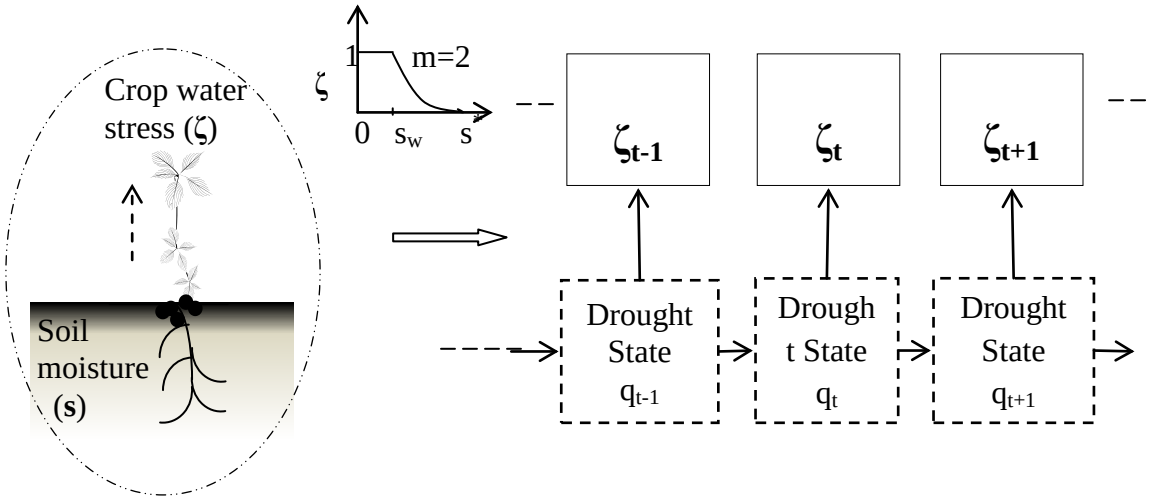

Figure 4.3 A schematic of the HMM used in this study. The hydrologic variable ζ_t represents the crop water stress. The hidden drought state q_t represents one of near normal, moderate, severe or extreme drought states. The subscript t is the time index. ζ is estimated from soil moisture content values s, s_w (at wilting point) and $s*$ (at incipient stomatal closure), and m is the measure of non-linearity in the estimation of ζ_t.

4.4.2 Temporal Dependence in Drought States

In the realm of statistical models, hidden Markov models are suitable for cases where temporal dependence in the drought states needs to be preserved. Otherwise, mixture models would suffice as a simpler tool for probabilistic modeling [Mallya et al. 2013a]. Mutual Information (MI) statistic is used in this study to determine the nature of temporal dependence between drought states at one-month interval. The drought states are based on a standardized crop-drought index calculated using the crop stress function values.

Mallya et al. [2013a] provide a detailed analysis of the nature of temporal dependence between drought states for meteorological and hydrological droughts with durations greater than one month in which case use of a hidden Markov model was favored, and highlighted the merits of adopting a simpler Gaussian mixture model (GMM) when temporal dependence was insignificant. However, for soil moisture-driven droughts, Markovian dependence in time cannot be neglected without exploring the nature of dependence, as soil moisture holds a long-term persistent memory [Manabe and Delworth, 1990; Koster and Suarez, 2001]. This aspect is investigated later in the chapter.

4.4.3 Graphical Models

A graphical model is a family of distributions that can be efficiently represented by a directed or undirected graph. Variables of interest are denoted by nodes whereas their dependencies are indicated by connections/edges. The graph structure allows users to compute marginal and joint conditional probabilities between variables present as nodes in the graph [Jordan, 2004]. Graphical models have been popular in the fields of speech recognition, language processing, genetics, and information retrieval; recent applications include modeling spatial and temporal patterns of precipitation [Ihler et al., 2007], and extreme event modeling [Yu et al., 2012].

4.4.3.1 Hidden Markov Models

Hidden Markov models are a class of graphical models where the graph structure comprises of hidden nodes with connections to observed nodes, such that temporal dependencies exist between the hidden nodes. In an HMM, as shown in Figure 4.3, the outputs/observations of the system are assumed to be dependent on a sequence of hidden

states. In the context of drought studies, the hidden nodes are the latent drought states, while the observations may be precipitation or streamflow values [Mallya et al., 2013a], or soil moisture-driven crop stress function as in this study.

Consider the model where the hydro-climatic variable of interest at an instant t is denoted by x_t, $t = 1, 2, \ldots, N$ $\{x_t \in R$ and $X = [x_1, x_2, \ldots, x_N]^T = x_{1:N}\}$. The observation x_t is dependent on the hidden state variable $q_t, \{Q = [q_1, q_2, \ldots, q_N]^T = q_{1:N}\}$ which is assumed to be a first order Markov process, i.e. the probability of the system being in any future step is independent of past states given the present state. The hidden variable q_t is considered a discrete variable representing one of the K possible latent states. The major characteristics of an HMM with K states and following first order Markov property are:

i. Given the state of the system at time $t-1$, q_t is independent of previous states i.e. $P(q_t \mid q_{t-1}, q_{t-2}, \ldots, q_1) = P(q_t \mid q_{t-1})$. The state transition probability matrix can be defined as $A = \{a_{ij}\}$ where $a_{ij} = P(q_{t+1} = j \mid q_t = i), \quad 1 \le i, j \le K$.

ii. Given the current state q_t, the observation at that instant x_t is conditionally independent of past observations, and the probability $P(x_t \mid q_t)$ is known as the emission distribution. The matrix $B = \{\alpha_i, \beta_i\}$ represents the parameters of the emission distribution.

iii. The initial state distribution, i.e., the probability that the drought state at the instant $t = 1$ $P(q_1)$ is given by $\pi = \{\pi_i\}$ s.t. $\pi_i = P(q_1 = i), \quad 1 \le i \le K$.

Besides, the following constraints hold valid for a HMM model:

$$\sum_{i=1}^{K} \pi_i = 1$$
$$\sum_{j=1}^{K} a_{ij} = 1; \quad 1 \le i \le K \tag{4.2}$$

That is, sum of the initial state probabilities and transition state probabilities respectively is equal to one. The joint distribution of the different drought states and observations in the HMM can then be expressed as

$$P(q_{1:N}, x_{1:N}) = \pi_i \prod_{t=2}^{N} P(q_t \mid q_{t-1}) \prod_{t=1}^{N} P(x_t \mid q_t) \tag{4.3}$$

4.4.4 Model Implementation

4.4.4.1 Emission Distribution

Gaussian emission distributions have been favored in several continuous-HMM applications due to ease of computation. However, there are applications where Gaussian densities cannot be used, and hence, parameter estimation methods have to be designed from first principles. A beta emission distribution was adopted in this study for the following reasons: (i) it is a continuous distribution, (ii) it is well-suited for variates over the finite range of [0,1], (iii) it has the flexibility to model very skewed emission distributions that are needed for extreme events, and (iv) distributional parameters can be estimated in the HMM context.

4.4.4.2 Parameter Estimation

An important task in generating a HMM-based probabilistic model for drought data is parameter estimation—finding the best set of $\{\pi, A, B\}$ such that the probability of the observation sequence given the model i.e., $P(O \mid \text{model})$ is maximized. Parameter

estimation in HMMs was performed using Baum-Welch algorithm that uses Expectation-Maximization [Baum et al., 1970; Rabiner, 1989]. The Baum Welch algorithm treats parameter estimation as a constrained optimization of $P(O|\text{model})$ subject to constraints in Equation (4.2), and estimation formulae for $\{\pi, A, B\}$ are developed using a Lagrange multiplier technique such that the results yield maximum $P(O|\text{model})$ value. The details of parameter estimation including that for the shape parameters $\{\alpha, \beta\}$ of the emission distribution are provided in Appendix A.

The initial user-input values fed into the HMM framework play an important role in the estimation of probabilities and parameter values as the estimation algorithm may run into local maxima during the simulations. In order to ensure global optima are achieved, random sets of initial values were tried, and the estimated values corresponding to maximum probability $P(O|\text{model})$ were chosen for the model. Thus parameter estimation was a trial and error method. In scaled HMMs, the term $\log\left[P(O|\text{model})\right]$ is maximized [Rabiner, 1989].

Once the model parameters are estimated, the conditional probability of being in a particular drought state at time t, given the observations and set of model parameters is simply the posterior probability of falling in that state at time t (see Appendix A, equation (A.10)). Probabilistic classification of drought states based on proposed crop water stress index is facilitated by estimating these probabilities using the HMM.

4.5 Results and Discussion

4.5.1 Crop Moisture Stress Estimation

Gridded soil moisture data at 52 locations over Indiana are used to compute the respective crop stress function values. Land cover data for these locations are retrieved from CDL provided by USDA-NASS. Only the major crops such as corn, soybean, sorghum, alfalfa, winter wheat, and double crops-winter wheat/soybean (WS) and winter wheat/corn (WC) are considered in the drought analysis. The average acreage distribution of various crops grown in Indiana is as follows: 35% to 55% each of corn and soybean, less than 10% each of winter wheat and double crop WS, and less than 1% of sorghum, alfalfa and WC.

For all these crops, the water requirements over their growing seasons are assessed based on rooting depths at different growth stages [Evans et al., 1996]. The adopted rooting depth variation with crop type and time of the year is shown in Table 4.1. Plant rooting depths were obtained mostly from past literature [Weaver and Bruner, 1927; Weaver, 1926; Rhoads and Yonts, 1991]. Soil water content s_w at permanent wilting point (PWP) and $s*$ at incipient stomatal closure required for crop water stress calculation are computed as percentages of water available in the root zone of the crops [Tolk, 2003], and these values are allowed to vary with different stages of plant growth. For instance, studies by Tolk [2003] determined PWP for corn and sorghum planted in 2-m deep soil to be around 488 mm and 420 mm respectively. For the different crops: soybean, alfalfa, and winter wheat, PWP, as a percentage of rooting depth are assumed to be 15, 10 and 19 percent respectively. The calculated monthly s_w and $s*$ values (in mm)

for different crops are shown in Figure 4.4. There is an increase in plant water requirement as the growth stage advances. These values are estimated based on the rooting depth values in Table 4.1 and plant water requirements mentioned previously. Under double cropping, values for s_w and $s*$ throughout the year are significant, unlike the case of a single crop as shown in Figure 4.4. The crop stress function time series is computed for the growing season of crops. A weighted crop stress function time series is then calculated using crop acreage data at each grid location.

Table 4.1 Rooting depths (in metres) for crops grown in Indiana over the annual growing season, where symbol '×' represents absence of cultivation [Weaver, 1926; Weaver and Bruner, 1927; Rhoads and Yonts, 1991].

Crop	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Corn	×	×	×	×	0.65	0.9	0.9	0.9	0.9	1	1	×
Soybean	×	×	×	×	0.5	0.76	0.9	1	1.4	1.5	1.8	×
Sorghum	×	×	×	×		1.2	1.55	1.65	1.85	1.85	×	×
Alfalfa	×	×	0.13	0.5	0.9	1	1.2	1.5	2	×	×	×
Winter wheat	0.8	1	1.2	1.3	1.3	×	×	×	×	×	0.5	0.6
WS*	0.8	1	1.2	1.3	1.3	0.5	0.9	1	1.5	1.5	0.5	0.6
WC[#]	0.8	1	1.2	1.3	1.3	0.65	0.9	0.9	1	1	0.5	0.6

*WS is double cropping, winter wheat + soybean
[#]WC is double cropping, winter wheat + corn

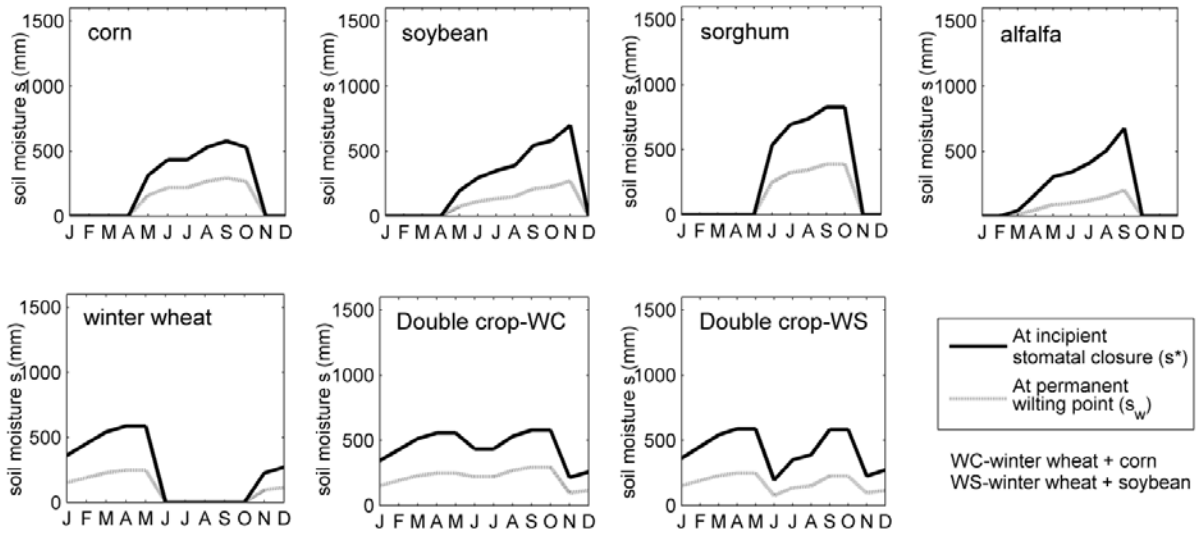

Figure 4.4 Monthly soil moisture content values at wilting point (s_w), and at incipient stomatal closure ($s*$) for various crops in the study region calculated based on crop growth stage and water requirements

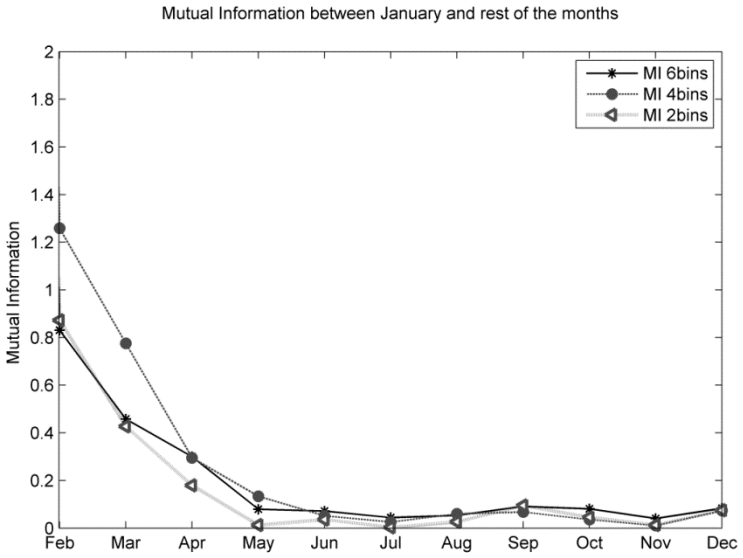

Figure 4.5 Mutual information statistic between standardized crop stress function values of January and rest of the months for 2, 4 and 6 bins

4.5.2 Exploring Temporal Dependence between Drought States

Figure 4.5 demonstrates the results of temporal dependence analysis conducted using mutual information statistic [MI; Cover and Thomas, 1991], where crop stress-based drought states for the month of January (as an example) are compared with those of other months. The crop stress function values are standardized and categorized similar to SPI-based drought classification (ranging from W4-W0, to normal to D0-D4; McKee et al., 1993]. For instance, in a two bin case, W4-Normal and D0-D4 classes are grouped into two drought states: no-drought and drought respectively. In a similar fashion, the categories are grouped into 4 and 6 bins for estimating temporal dependence. For each of these cases, respective monthly MI statistics were computed using Equation (4.4).

$$MI(X,Y) = \sum_{x \in \mathrm{X}} \sum_{y \in \mathrm{Y}} p_{x,y}(x,y) \log \frac{p_{x,y}(x,y)}{p_x(x)\, p_y(y)} \tag{4.4}$$

where $p_{x,y}, p_x, p_y$ are joint probability of (X,Y), and marginal probabilities of X and Y respectively. As an example, mutual information statistic values between drought states in January (X) and those in the rest of the months of the year (Y) were calculated from monthly time series of ζ for one station, and are plotted in Figure 4.5. It is seen that the January drought states share temporal dependence with those of February and March, based on higher MI statistic values. The conclusion was same from results at other locations, and for other months, i.e. temporal dependence among drought states cannot be ignored. To account for the dependence in drought states while modeling even one-month droughts, HMMs are needed over the simpler mixture models.

As the number of hidden states increases in HMMs, the corresponding number of model parameters also increases, adding significantly to model complexity and data requirements. In the present study, HMMs with four hidden states are considered for probabilistic assessment of agricultural droughts. These hidden states would represent instances of near normal conditions, moderate, severe and extreme droughts, respectively. Further, as soil moisture changes slowly, the transition probabilities are modeled by a tridiagonal matrix, implying the system could continue in the present state or move to a one-level drier or wetter state over a single time step. These model constraints ensured smaller number of model parameters and more stable results.

4.5.3 Development of Probabilistic Drought Model

An HMM-based probabilistic drought classification was developed using the crop water stress values at all the 52 grid locations in Indiana as the drought states do share dependence in time. The parameter estimation procedure included initialization and estimation of initial state probabilities, transition probability matrix and beta emission distribution parameters. Scaled HMM [Rabiner, 1989] was used herein to facilitate parameter estimation. The best set of parameters was identified based on maximum $\log\left[P(O\,|\,\text{model})\right]$ value from the simulations obtained from random initial values. As noted earlier, a tridiagonal transition matrix was assumed at the second stage of parameter estimation, after having set the order of hidden states in increasing order of drought severity. The best parameter values were then obtained from simulations using random values as transition probabilities, with previously estimated beta emission

parameters to initialize the new α and β parameters of beta-HMM. Parameter estimates for the HMM model at six locations in Indiana are shown in Table 4.2 and Figure 4.6 as representative samples, for the sake of brevity. These are geographically widely separated points and are denoted by their location identifiers (loc id): 7 (41.25°N, 87.25°W), 9 (41.25°N, 86.25°W), 12 (41.25°N, 84.75°W), 46 (38.25°N, 87.25°W), 35 (39.25°N, 85.75°W), and 44 (38.75°N, 84.75°W) respectively. As expected, in most cases, preference is expressed for continuing in the present state than transitioning to a neighboring state.

The emission distributions in Figure 4.6 allow for some statistical interpretation into the drought states. They represent the changing nature of agricultural droughts with spatial locations. At all locations, the emission distribution for near normal conditions have very peaked distributions with a large probability mass concentrated close to ζ=0. At loc. id 7, as seen in Figure 4.6a, the emission distributions for all drought classes have reasonable separation implying that the model is able to resolve these classes with less uncertainty. The peaked probability density functions for near normal and extreme drought states at all locations indicate that these categories are classified with higher probabilities. However, for loc. id 7 and 12, high classification uncertainty exists for severe and extreme droughts (Figure 4.6, plots a, c), as the emission distributions have more overlap for severe and extreme drought classes. Consequently, higher transition probabilities exist for transition of extreme drought to severe drought state at these two locations (Table 4.2 a, c). The moderate drought class, on the other hand has very little overlap in all the six cases, implying less uncertainty in its classification.

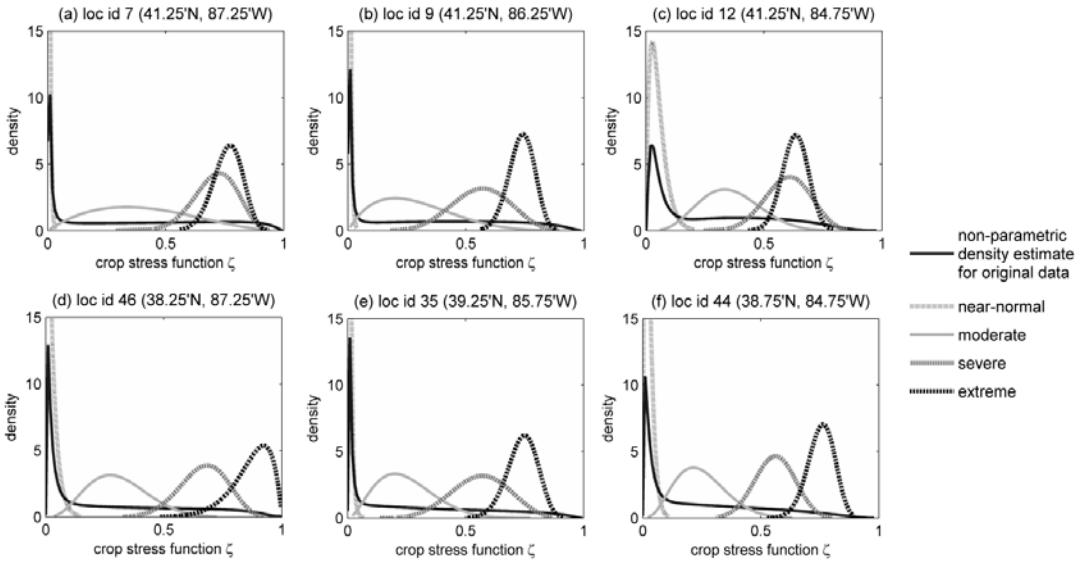

Figure 4.6 Estimated emission densities (beta distribution probability density functions) for six locations across Indiana

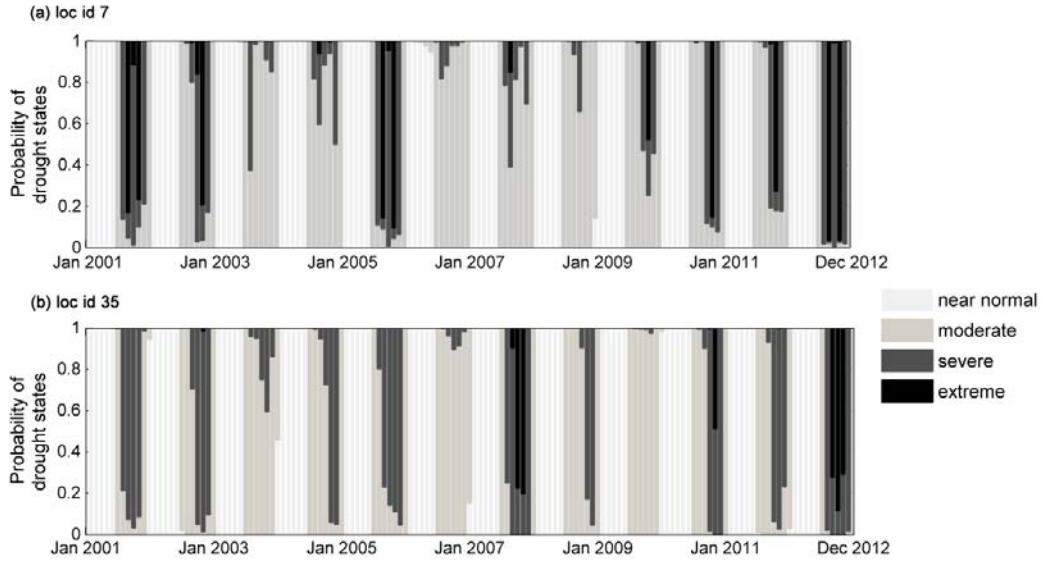

Figure 4.7 Probabilistic classification of agricultural droughts during 2001-2012 period at (a) loc id 7 and (b) loc id 35 using the proposed crop stress-based index

Table 4.2 Estimated hidden Markov model probabilities- initial state (π_i) and transition state probabilities, and beta emission distribution parameters α and β associated with the four drought states (1-near normal, 2-moderate, 3-severe and 4-extreme) for six locations in Indiana

		(a) loc id 7				(b) loc id 9				(c) loc id 12			
Drought State ➔		1	2	3	4	1	2	3	4	1	2	3	4
	π_i	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Transition Probabilities	1	0.80	0.20	0.00	0.00	0.80	0.20	0.00	0.00	0.81	0.19	0.00	0.00
	2	0.21	0.62	0.17	0.00	0.34	0.27	0.39	0.00	0.23	0.60	0.18	0.00
	3	0.00	0.54	0.01	0.45	0.00	0.34	0.57	0.09	0.00	0.48	0.01	0.51
	4	0.00	0.00	1.00	0.00	0.00	0.00	0.53	0.47	0.00	0.00	1.00	0.00
	α	1	2	17	35	1	2	9	47	2	5	15	48
	β	221	3	7	11	188	5	7	17	37	9	10	28

		(d) loc id 46				(e) loc id 35				(f) loc id 44			
Drought State ➔		1	2	3	4	1	2	3	4	1	2	3	4
	π_i	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Transition Probabilities	1	0.81	0.19	0.00	0.00	0.80	0.20	0.00	0.00	0.80	0.20	0.00	0.00
	2	0.25	0.55	0.20	0.00	0.27	0.47	0.26	0.00	0.22	0.60	0.18	0.00
	3	0.00	0.32	0.67	0.01	0.00	0.34	0.59	0.07	0.00	0.33	0.63	0.05
	4	0.00	0.00	0.51	0.49	0.00	0.00	0.57	0.43	0.00	0.00	0.46	0.54
	α	1	4	14	13	1	3	9	34	2	4	19	43
	β	41	9	7	2	130	9	7	12	76	12	15	14

Figures 4.7a and b show probabilistic classification of drought states provided by the crop water stress-based drought index in HMM framework. Results from only two locations for an example 12 year period 2001-2012 are shown here for the sake of brevity. The height of each bar in the plots represents the probability of a particular drought state in a particular month. While the lighter shade represents a near-normal condition, the darker ones represent increasing severity of drought induced by crop water stress. For instance in Figure 4.7a, July 2012 at loc id 7 had the following drought

probabilities: 98.2% of being in severe drought, and 1.6% and 0.2% of being in moderate and extreme states respectively. Similarly, HMM-based classification for August 2012 at loc id 35 indicates 72.5% and 27.5% probabilities of being in extreme and severe states respectively (Figure 4.7b). In contrast to popular indices such as SPI, SPEI and PDSI, the probabilistic drought state classification offered by the proposed index addresses uncertainty in drought characterization. Comparisons with these indices are discussed in the following section.

4.5.4 Comparison with Popular Drought Indices

Most drought studies have relied on the PDSI (based on a soil water balance equation), and the SPI (based on a precipitation time series). Instead of PDSI, a self-calibrating PDSI (SC-PDSI) that can account for the regional variability in climate [Wells et al., 2004] was used for comparison purposes. As the PDSI is not multiscalar, and a fully meteorological-based SPI cannot provide any indication of crop water stress, both these indices are incapable of evaluating agricultural droughts at different locations in Indiana. SPEI-based analyses conducted by Vicente-Serrano et al. [2012] show that SPEI possesses good correlation with soil moisture in most of the sites in North America. The SPEI computation uses monthly precipitation minus potential evapotranspiration, i.e. a water balance deficit data series, that is aggregated at different time scales as in SPI [McKee et al. 1993], and standardized using a three-parameter log-logistic distribution [Vicente-Serrano et al., 2010]. SPEI time series were computed using SPEI calculator program developed by Beguería and Vicente Serrano [2009]; inputs for the program include precipitation and temperature data, as well as the latitude of the selected location.

Therefore, for comparison purposes, the SPEI index is also utilized, and relative merits and demerits of all the four indices are evaluated.

Drought category classifications for all indices used in the study are listed in Table 4.3. Unlike SPI and SC-PDSI, drought categorization with SPEI is fairly recent [Yu et al., 2013]. For the proposed HMM-based index, there is no hard classification, and the probability associated with each drought state at a given time can be obtained. For comparison purposes, the predominance of a particular state is indicated when the probability of falling in it exceeds the sum of probabilities of falling in the other states.

Table 4.3 Drought category classification of the common drought indices

Hidden State	Drought Definition	SPI (McKee et al., 1993)	SPEI (Yu et al., 2013)	SC-PDSI (Wells et al., 2004)
1	Near normal	+1 to -0.99	+1 to -0.99	+0.5 to -0.99
2	Moderate drought	-1 to -1.49	-1 to -1.49	-1 to -2.99
3	Severe drought	-1.5 to -1.99	-1.5 to -1.99	-3 to -3.99
4	Extreme drought	Less than -2	Less than -2	Less than -4

Figures 4.8 and 4.9 show the probabilistic monthly drought classification offered by HMM and the corresponding SPEI, SC-PDSI and SPI index values during an example 20 year period - from 1983 to 2003 at loc. id 7 and 35 respectively. The HMM-based method yields probabilities associated with each drought category, thus providing a basis for assessing classification uncertainty, unlike SPI, SPEI or SC-PDSI. At loc. id 7, (Figure 4.8), few extreme and severe agricultural drought events are identified in the years 1983-1985, 1988, 1995, 1999-2002, according to the proposed crop stress-based index. SPI and SPEI reported extreme droughts in 1984, 1988, 1991-1992, 1999-2000.

On the other hand, SC-PDSI detected very few extreme events during this period, in 1985-1986 and 1993. Severe droughts according to the SPI and SPEI indices, occurred in 1985-1987, 1992 and 2000, and are identified by the proposed index as well. All the indices suggest that near normal to moderate drought conditions are more prevalent in loc id 7. In Figure 4.9, at loc. id 35, very few extreme events are suggested in 1988 and 1999 by the proposed index, and severe drought events are more prevalent. SPI and SPEI projected extreme droughts for years 1988, 1991, and 2002, whereas SC-PDSI reported extremes in 1992-1993 and 2003. Moderate drought events are observed frequently during June-September months. The results at these two locations therefore suggest that the developed probabilistic index is capable of identifying agricultural drought events that may not be captured by the SPI, SPEI or SC-PDSI, especially during the months of May-October, the growing season for most of the crops. Additionally, the probabilities assigned to each drought category in the HMM-based probabilistic classification reflect the uncertainty involved in drought identification. The other indices were not designed for this capability.

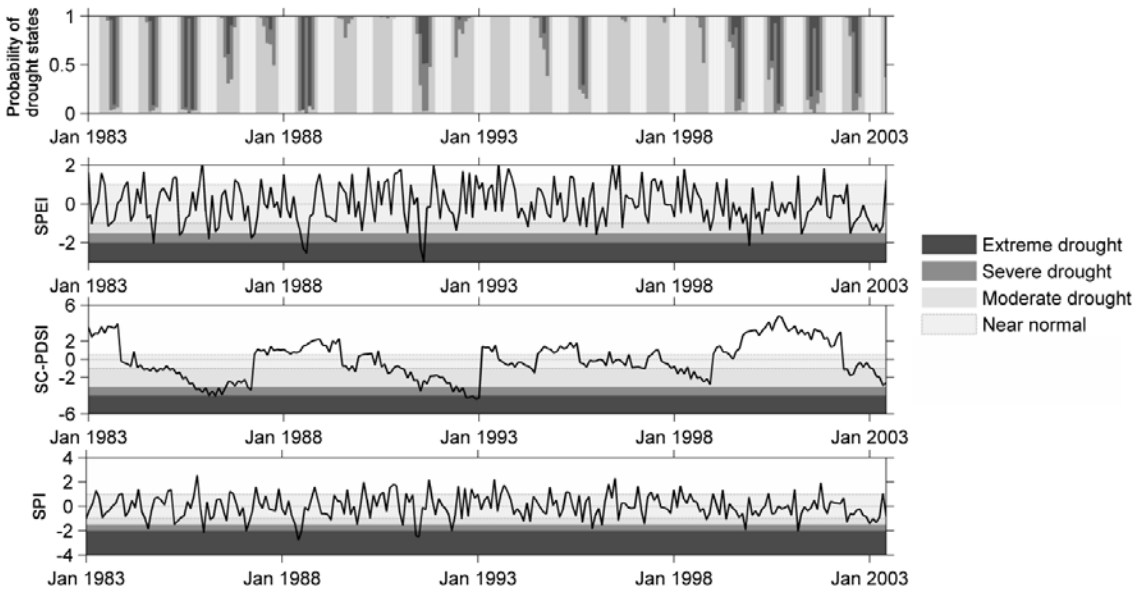

Figure 4.8 Comparison between HMM-based agricultural drought index, SPEI, SC-PDSI and SPI values for location id 7 (lat/lon 41.25°,-87.25°) during the 1983-2003 period

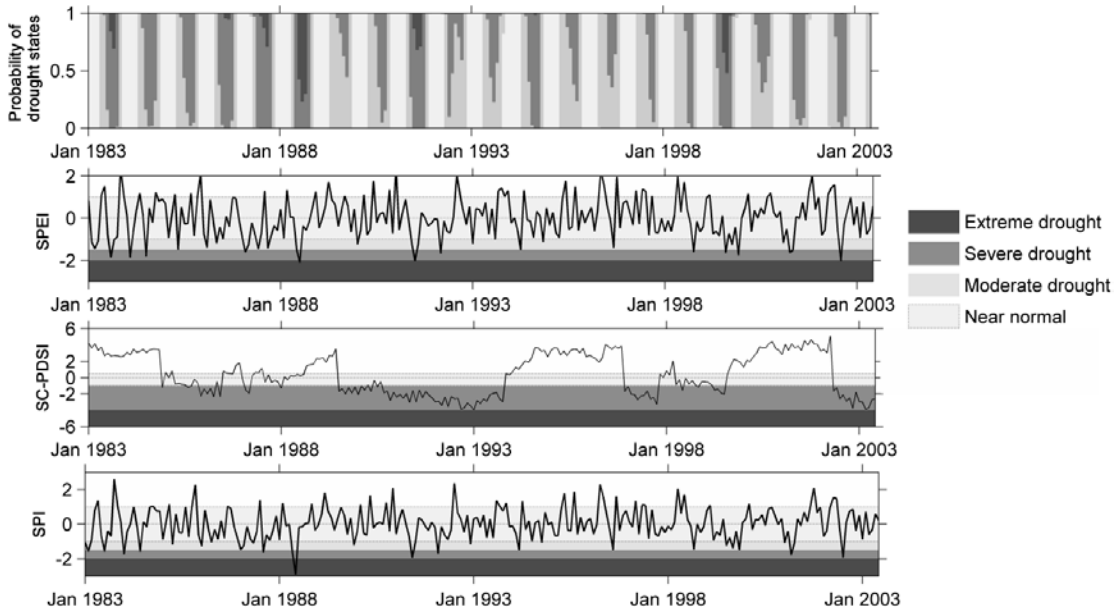

Figure 4.9 Comparison between HMM-based agricultural drought index, SPEI, SC-PDSI and SPI values for location id 35 (lat/lon 39.25°,-85.75°) during the 1983-2003 period

Since different indices are designed for different purposes and yield different information, the superiority of any one index over others cannot be established. Comparisons between results from different indices may imply robustness if results are consistent. For example, the number of extreme events detected by the proposed index and SC-PDSI during the data period 1948-2012 in Indiana is shown in Figure 4.10, pooling drought information from across all the 52 locations in Indiana. Darker shades correspond to increased frequency of extreme droughts during 1948-2012. According to the proposed crop stress-based index, northern Indiana is relatively more prone to extreme agricultural droughts, while southwest Indiana has had relatively few instances over the data period. The drought maps for extreme events from the proposed index and SC-PDSI are markedly different, suggesting that different indices may lead to different conclusions. There is some agreement in the extreme drought occurrences suggested by the proposed index and SC-PDSI for south-eastern, south-central and central Indiana, but the proposed index would suggest that the state is more prone to extreme droughts.

Similarly, severe drought event maps were constructed for Indiana using the two indices and are shown in Figure 4.11. The ranges of number of severe events during the period 1948-2012 identified by the proposed index and SC-PDSI are vastly different. The proposed index reported numerous instances of severe droughts all over the region, far more than those identified by SC-PDSI. The SC-PDSI maps in Figures 4.10 and 4.11 consistently indicate that west and central Indiana have experienced high frequency of extreme and severe category droughts over the 1948-2012 period. However, the proposed index suggests that central and southern Indiana are highly prone to severe droughts.

However, it has to be noted that one index is not superior to the other, just that different indices may yield different results implying the choice of an index for drought classification should be based on the specific needs of the user. An evaluation of relative drought-proneness of a region cannot be evaluated by SPI and SPEI as all locations are allocated the same probability of a drought class by definition.

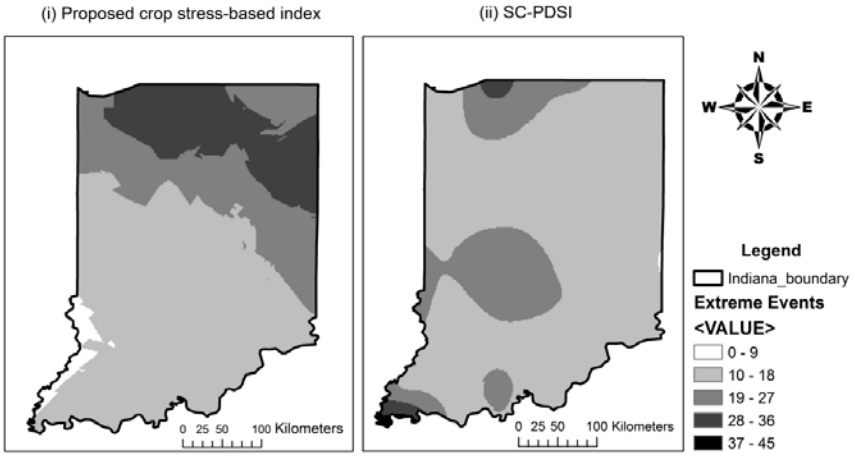

Figure 4.10 Extreme drought category maps for Indiana under (i) the proposed crop stress-based index, and (ii) SC-PDSI

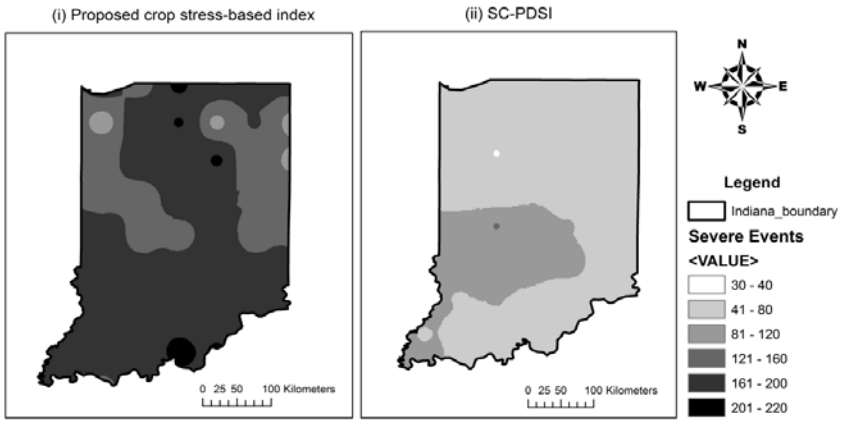

Figure 4.11 Severe drought category maps for Indiana under (i) the proposed crop stress-based index, and (ii) SC-PDSI. SC-PDSI reports a smaller range of occurrences compared to the proposed index

4.6 Summary and Conclusions

A probabilistic agricultural drought index that is based on crop water needs was formulated within a graphical model (HMM) framework, where hidden states represent different drought categories (from near normal to extreme droughts). The monthly soil moisture-based crop water stress function estimated in the study was found to have temporal dependence between drought states, thus suggesting the use of HMMs over simpler mixture models. Crop water stress was modeled using HMMs with a tridiagonal transition matrix and beta emission densities to develop a probabilistic model based on a bounded stress function.

Retrospective comparison of drought events of an example 20 year period (1983-2003) suggested by the proposed model and indices such as SPEI, SC-PDSI and SPI indicated fairly good agreement over agricultural drought conditions. Given that consistent definitions of corresponding SPI, SPEI and SC-PDSI index values for each drought state in the HMM framework−near normal, moderate, severe and extreme droughts are not available, direct comparisons could not be made. Focusing on the crop stress-based index for the 1988 and 2012 droughts at loc id 7 and 35, its severe and extreme category droughts were identified with very high probabilities by the index from as early as the summer of 1988, and their persistence was observed for a longer time, i.e. 5 to 6 months (Figures 4.8, 4.9). The other indices−SPEI and SPI, indicated similar drought magnitudes for certain months of the year, however, drought withdrawal was relatively early. Similarly, the 2012 drought period, though not shown in the figures, was dominated by high probability of severe and extreme events at loc id 7, and 35,

respectively. Early onset of droughts and longer persistence are suggested by the proposed index when compared to the popular indices. Additionally, extreme and severe drought category maps were developed for whole of Indiana using results from the proposed crop water stress-based index, to study the spatial variation of drought-proneness of the study region.

The following observations are made regarding the probabilistic agricultural drought index developed in this chapter:

i. Drought severity category is defined differently for each location by the HMM. Drought states evolve based on the historical crop water stress time series at each location, and hence, an averaged or aggregated assessment for a region cannot be considered accurate.

ii. The tridiagonal transition matrix assumption adopted in HMMs in this study holds good for smooth transitioning of drought states and facilitates robust parameter estimation. However, sudden drought transitions that occur in the case of flash droughts may not be well captured by the model under this assumption.

iii. The transition trends and emission distributions are not similar over Indiana. Results tend to be site-specific, suggesting the need for advanced regionalization studies for regional agricultural drought outlook.

iv. For comparisons with existing drought indices, the predominant drought category after probabilistic classification was defined as the one whose probability of occurrence was more than the sum of probabilities of droughts in all other categories.

v. In the event that no drought category is dominant, the classification uncertainty is likely to be high, i.e. multiple drought categories are about equally likely. In the present study, predominant drought categories were distinctly identified over the study area.

vi. Comparison of indices indicated that many drought events during dominant crop growing season (May-October) that were not identified by the SPI, SPEI and SC-PDSI, were revealed by the proposed index.

vii. The spatial variation of propensity of extreme and severe category droughts over Indiana during the 1948-2012 period was examined by the proposed crop stress-based index (Figures 4.10 and 4.11). Such maps are useful for planning crop cultivation under rain-fed conditions. Since different indices yield different results, the choice of the index should be based on the desired end result. The utility of these maps need to be further explored in identifying regions where certain crops can be cultivated with minimum chances of crop water stress.

The proposed HMM-based drought index enables classifying agricultural droughts in a probabilistic framework unlike the SPI, SC-PDSI or SPEI. The graphical model-based index highlights the inherent uncertainty in drought analysis, and the framework would be useful in developing reliable forecasting models. The crop water stress-based drought index developed using HMMs also suggests the need for alternate drought classification regimes that are driven by the data.

The current study has not accounted for tile drain systems that are prevalent in agricultural fields in Indiana. The tiles that are laid at the level of water table (usually 2-4 feet below the surface) serve as a boundary for root growth. The crop rooting depths need to match the field conditions in such locations. The rooting depths are therefore lower than those currently used, from Table 4.1, and therefore, crop stress values could be lower than the current modeled values.

Another important factor to consider is the varying water demand of crops with the growing season. It was assumed that the growing season was as shown in Table 4.1. The root depths need to be better assessed for the crops being grown at a particular location depending on their growing stage, and the data used in Table 4.1 can be fine-tuned for local conditions.

The uncertainty involved with the modeled soil moisture data used in the study has not been accounted for in the model results. Observed soil moisture data could be used to avoid data discrepancies. For instance, in few locations in Indiana, soil moisture sensors are installed to collect soil moisture and related data round the year (http://amarillo.nserl.purdue.edu/ceap/index.php). However, these data sets are not sufficiently long, have coarse spatial resolution, and the sensor locations are not stable.

CHAPTER 5. CHOICE OF HYDROLOGIC VARIABLES FOR PROBABILISTIC DROUGHT CLASSIFICATION: A CASE STUDY

5.1 Abstract

Watershed-scale drought assessment is performed using cumulative density function (CDF)-based probabilistic drought indices in this study. To investigate the role of hydrologic variables, in combination, copulas are used for multivariate joint cumulative density functions (CDFs) combined with graphical models for probabilistic drought classification. Adopting a multivariable, multiscalar approach in the proposed framework yields a drought index that allows for examining the roles of hydrologic variables on integrated drought assessment. The methodology is demonstrated using streamflow, precipitation and soil moisture anomalies to develop univariate and multivariate CDF-based indices at 1-, 3- and 6-month time scales to analyze the drought events over an Indiana watershed. Drought characterization varied across the univariate, bivariate and trivariate drought models in the case study. The multivariate models were able to capture the early onset of drought events and persistence of the drought states, features that are contributed by different components of the hydrologic cycle. While short term drought monitoring is facilitated by 1-month models, threats to long term water-storage in the watershed can be assessed better with longer time scale models.

5.2 Introduction

Drought characterization using individual hydro-climatic variables is very popular within the hydrologic community, and there are drought indices that specifically cater to meteorological, hydrological and agricultural drought studies. Indices such as SPI [McKee et al., 1993], crop moisture index [CMI; Palmer, 1968], standardized runoff index [SRI; Shukla and Wood, 2008], and surface water supply index [SWSI; Shafer and Dezman, 1982], are few examples. The onset, severity and duration of droughts detected by the use of different hydrological variables may vary, and overall drought assessment is often performed by combining various hydrologic variables or by performing multivariate analyses. Drought studies that deviate from the standard univariate drought classification scheme advocate that (i) a single variable-based analysis may not be sufficient to address the overall drought condition at a location, and (ii) dependencies between hydro-meteorological variables leading to droughts should be utilized to characterize droughts in a better fashion. Indices such as the PDSI [Palmer, 1965], aggregate drought index [ADI; Keyantash and Dracup, 2004], hybrid drought index [HDI; Karamouz et al., 2009], standardized precipitation evapotranspiration index [SPEI; Vicente-Serrano et al., 2010], multivariate standardized drought index [MSDI; Hao and AghaKouchak, 2013,2014), joint drought index [JDI; Kao and Govindaraju, 2010], and the United States drought monitor [USDM; Svoboda et al., 2002], utilize information from multiple drought indicators for drought classification.

The PDSI and SPEI are based on water balance deficit computed using observed precipitation and precipitation, temperature and the local available water content (AWC) of the soil as inputs, respectively, however, neither of them account for streamflows. The ADI index uses the standardized first principal component of six different variables (precipitation, evapotranspiration, streamflows, reservoir storage, soil moisture, snow water content) to encompass the influence of multiple hydrologic variables on drought classification. Principal components (PCs), while honoring variability in the data, do not allow for physical interpretation. Recently, Rajsekhar et al. [2014] developed a multiscalar multivariate drought index (MDI) that utilized SPEI, SRI, and standardized soil moisture index [SMI; Hao and Aghakouchak, 2014] as inputs to account for meteorological, hydrological and agricultural droughts, respectively. The MDI was formulated using kernel entropy component analysis (KECA) to preserve the maximum amount of information from the input drought indicators.

For bivariate and multivariate joint formulations, copulas are used for scale-free association between different variables irrespective of their marginals. The popularity of copulas has grown from financial and insurance models to meteorology and hydrology in the last two decades [e.g. Salvadori and De Michele, 2004; Grimaldi and Serinaldi, 2006; Favre et al. 2004; Zhang and Singh, 2006; Shiau, 2006; Kao and Govindaraju, 2008; Maity et al., 2013]. Shiau [2006] used the SPI to define droughts, and the marginals of drought duration and severity were used in copula framework to construct the joint distribution. Serinaldi et al. [2009] used a four dimensional student copula to model SPI drought properties namely the duration, mean and minimum SPI values, and drought mean areal extent, and to compute the joint return periods and exceedance probabilities.

Kao and Govindaraju [2010] used bivariate copulas of precipitation and streamflows to define the joint drought index (JDI). The MSDI index proposed by Hao and AghaKouchak [2013] for overall characterization of droughts was based on a joint dependence model of SPI and standardized soil moisture index (SSI) using bivariate Frank and Gumbel copulas. MSDI captured early onset of precipitation-driven droughts as well as delayed persistence of soil moisture-driven droughts.

Steinemann [2003] had proposed a cumulative density function (CDF) or percentile-based index for developing, comparing and evaluating drought precursors as it provides a consistent basis for comparing multiple drought indicators. It was argued that percentiles are statistically comparable across spatial and temporal scales, irrespective of the drought indicator variables used in the study. The author suggested classifying the percentiles using thresholds for different drought categories ranging from 1 to 6 in increasing order of drought severity. The classification thresholds were {1, 0.50, 0.35, 0.20, 0.10, 0.05, 0}.

While previous drought studies have used hydrologic variables either singly or in combination, a question that has received little attention is the relative role of these hydrologic variables in drought classification. For instance, how does drought characterization change with different combinations of hydrologic variables? Are all variables needed for overall drought assessment, or would a smaller subset suffice? If so, what variables should be included in this smaller subset? Previous studies have not directly addressed these questions. The answers to these questions will change with location, study areas (watersheds), and perhaps how the indices are chosen.

The goal of this case study is to propose one method for understanding the role of hydrologic variables and answering the aforementioned questions. In order to assess the uncertainty in drought classification and preserve the temporal memory in drought states, graphical models, specifically hidden Markov models [HMMs; Rabiner, 1989], have shown promise [Mallya et al., 2013; Ramadas and Govindaraju, 2014]. These studies were based on a single variable, and the dimensionality of the HMM parameter space versus the length of available data was a crucial factor in robust parameter estimation. Probabilistic classification in a multivariate framework will result in a larger parameter space, aggravating the consequences of *curse of dimensionality*. Dimensionality reduction techniques such as principal component analysis (PCA) may be used, however, the PCs may not capture most of the variance in the non-Gaussian and dependent variable data used in drought analyses [Han and Liu, 2013]. Copulas are therefore used to combine drought-related variables to reduce dimensionality of the drought indicator in this study. The joint CDF of the hydrologic variables will yield a less complex HMM framework for multivariate drought models.

In this case study, probabilistic multiscalar drought indices were utilized using cumulative probabilities of marginals and joint distribution functions of anomalies of streamflows, precipitation and soil moisture as representatives of hydrological, meteorological and agricultural droughts, respectively, to address overall drought status of an Indiana watershed. Even with only three primary hydrologic variables, there are seven cases to consider–three univariate, three bivariate and one trivariate drought classification models are examined. In contrast to copula-based drought indices such as JDI [Kao and Govindaraju, 2010] and MSDI [Hao and AghaKouchak, 2013], the

cumulative probabilities from the joint CDFs were utilized to characterize droughts. The CDF value ranges from [0,1], and therefore, a beta emission HMM [Ramadas and Govindaraju 2014] was used for probabilistic drought categorization. Comparison of results from univariate and multivariate analyses shed light on the dependencies between the meteorological, hydrological and agricultural droughts. This allows assessment of the merits of using a multivariate index to assess drought status of the region. Additionally, HMM-based model accounts for uncertainty in state classification. The study further discusses the implications of the results at different time scales−1-month, 3-months and 6-months. The rest of the chapter is organized as follows: section 5.3 discusses the data used in the study, the methodology is elaborated in section 5.4, results and discussion of comparisons of indices follow in section 5.5, and conclusions from the study are presented in section 5.6. The model results for 3- and 6-month models are included in Appendix B.

5.3 Data Used in the Study

The study area is an agricultural watershed in the Ohio river basin, in Indiana, USA, and extends from 38°34’N to 39°49’N and 85°24’W to 86°31’W, covering an area of 6259 square kilometers. The watershed delineation was carried out using 30 m resolution digital elevation model (DEM) from USGS National Elevation Data set.

The drought-related variables used in the study are precipitation, soil moisture and streamflows, all at monthly time step. Modeling the dependencies of a drought requires a long record of historic observations, and 50-years minimum is recommended by previous

studies [Bonnin et al., 2004; Kao and Govindaraju 2010]. Precipitation and soil moisture values were obtained from the Climate Prediction Center (CPC) soil moisture model [Huang et al., 1996; Fan and van den Dool, 2004] for the period 1958-2012. While precipitation data are observed, soil moisture values were modeled by the 'leaky bucket' hydrological model of Huang et al. [1996] assuming a soil depth of 1600 mm, and the data are available for locations globally at 0.5° resolution and on a monthly time step. The watershed-scale drought study required spatially lumped data, and thiessen polygon method was used to compute the spatially averaged data set from the values at various grid points lying in the watershed.

The US Geological Survey (USGS) monthly streamflow data recorded at the USGS 03371500 (East Fork White River near Bedford, Indiana) from 1958-2012 were used in the present study. Hydrologic studies involving low flows have to ensure that the flows are not regulated, i.e., they are not influenced by any storage or release controls. Hence, drought analysis was carried out in an unregulated watershed in this study.

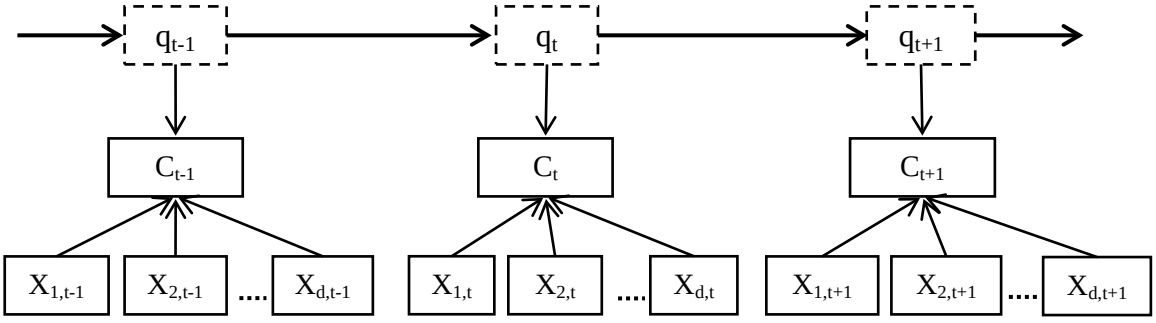

Figure 5.1 Schematic of multivariate (d-dimensional) drought classification scheme using a hidden Markov model (HMM). Here, X_1, X_2,…,X_d are the hydrologic variables used in the case study, C is the joint CDF or the joint probability distribution, and q is the hidden drought state. Subscript t stands for time step

5.4 Methodology

The schematic of a *d*-dimensional multivariate drought classification model at 1-month time scale is shown in Figure 5.1. The variables−streamflows, precipitation and soil moisture−and their different combinations are explored as drought indicators in a graphical model framework in this case study. The CDF of multivariate model of hydrologic variables (C) are generated using copulas. The various steps in the construction of models for drought monitoring are explained in this section.

5.4.1 Data Processing

Hydroclimatic variables−precipitation, soil moisture, and streamflows−at monthly time step were converted into anomalies by deducting the corresponding long term monthly mean from these variables. Let X_1, X_2, X_3, represent the variable anomalies of streamflow, precipitation, and soil moisture- the inputs to the multivariate drought model shown in Figure 5.1. Then, their marginal probabilities are denoted by $u_1 = F_1(x_1), u_2 = F_2(x_2),$ and $u_3 = F_3(x_3)$. These marginals are obtained by fitting suitable distributions to the variable anomaly data. The candidate distributions for variable anomalies were extreme value, generalized extreme value (GEV), normal, and student's *t* distributions.

5.4.2 Bivariate and Multivariate Copula Models

Copulas are defined as functions that join multivariate distributions to their one-dimensional marginals. Especially when the individual variables are non-normal, copulas offer a viable and straightforward alternative to modeling of different parametric families of distributions. According to Sklar [1959], a d-dimensional CDF with univariate margins $F_1, F_2, ..., F_d$ is defined by

$$H(x_1, x_2, .., x_d) = C(F_1(x_1), F_2(x_2), ..., F_d(x_d)) = C(u_1, u_2, ..., u_d) \tag{5.1}$$

where $F_k(x_k) = u_k$ for $k = 1, 2, ..., d$ with $U_k \in U(0,1)$ if F_k is continuous.

Hence, in the context of drought indicator variables, the bivariate copula of two variables X_1 and X_2, and trivariate copula of three variables X_1, X_2 and X_3, are, respectively, $C(F_1(x_1), F_2(x_2))$ and $C(F_1(x_1), F_2(x_2), F_3(x_3))$. Clayton, Gaussian, Frank, Gumbel, and student's *t* bivariate copulas were selected as candidates to model the joint behavior of pairs of these variables, and each of these are characterized by a single dependence parameter θ. For three dimensional joint distributions of variables, Gaussian and student's *t* copulas were explored. Additionally, fully nested or asymmetric Archimedean copulas were used to model the trivariate joint distributions. A *d*-dimensional nested copula is given by $d-1$ distinct generating functions as:

$$C(u_1, u_2, ..., u_d) = C_1\left(u_d, C_2\left(u_{d-1}, ..., C_{d-1}\left(u_2, u_1\right)...\right)\right) \tag{5.2}$$

There are $d(d-1)/2$ ways of coupling d variables in a multivariate model, as shown in Equation (5.2). A nested 3-copula model is characterized by two parameters, θ_1 and θ_2 such that $\theta_1 \le \theta_2$, such that higher degree of dependence exists between the inner

nested variables. Two dependence structures are present for three possible pairs in this case [Grimaldi and Serinaldi, 2006]. The bivariate and trivariate copulas along with their dependence parameters are listed in Table 5.1. Among the trivariate copulas, M3, M4, M5 and M6 families are the fully nested copulas, and further details of these families can be obtained from Joe [1997] and Embrechts et al. [2003]. Using maximum likelihood approach, copulas in Table 5.1 were fit to the multivariate data models to obtain parameter estimates. For detailed definitions and unique properties of copulas, as well as the parameter estimation procedures, the readers are requested to refer to previous studies [Maity et al., 2013]. For the sake of brevity, in this study, descriptions of two- and three-dimensional copula models, parameter estimation, and the best copula selection procedure, are limited to relevant details only.

Table 5.1 Bivariate and trivariate copula families selected for the study

Bivariate Families	
1	Clayton copula: $C(u_1,u_2;\theta)=(u_1^{-\theta}+u_2^{-\theta}-1)^{-1/\theta}\;;0\le\theta\le\infty$
2	Frank copula: $C(u_1,u_2;\theta)=-\theta^{-1}\log([(1-e^{-\theta})-(1-e^{\theta u_1})(1-e^{\theta u_2})]/(1-e^{-\theta}))\;;0\le\theta<\infty$
3	Gumbel copula: $C(u_1,u_2;\theta)=\exp\{-((-\log u_1)^{\theta}+(-\log u_2)^{\theta})^{1/\theta}\}\;;1\le\theta<\infty$
4	Gaussian copula: $C(u_1,u_2;\theta)=\Phi_{\theta}(\Phi^{-1}(u_1),\Phi^{-1}(u_2))\;;0\le\theta\le 1$ where Φ is the standard normal distribution N(0,1) with mean zero and unit variance, and Φ_{θ} is the bivariate standard normal distribution with correlation θ
5	Student's *t* copula: $C(u_1,u_2;\vartheta,\Sigma)=t_{\vartheta,\Sigma}(t_{\vartheta}^{-1}(u_1),t_{\vartheta}^{-1}(u_2))\;;1\le\vartheta<\infty;\Sigma\in\mathbb{R}^{m\times m}\;;$ where $t_{\vartheta,\Sigma}$ is student's *t* distribution with a correlation matrix Σ with ϑ degrees of freedom

Table 5.1 Bivariate and trivariate copula families selected for the study (continued)

Trivariate Families	
1	Gaussian copula: $C(u_1,u_2,u_3;\theta)=\Phi_\theta(\Phi^{-1}(u_1),\Phi^{-1}(u_2),\Phi^{-1}(u_3));0\le\theta\le 1$ where Φ is the standard normal distribution N(0,1) with mean zero and unit variance, and Φ_θ is the trivariate standard normal distribution with correlation matrix θ
2	Student's *t* copula: $C(u_1,u_2,u_3;\vartheta,\Sigma)=t_{\vartheta,\Sigma}(t_\vartheta^{-1}(u_1),t_\vartheta^{-1}(u_2),t_\vartheta^{-1}(u_3));1\le\vartheta<\infty;\Sigma\in\mathbb{R}^{m\times m}$ where $t_{\vartheta,\Sigma}$ is the student's t distribution with a correlation matrix Σ, with ϑ degrees of freedom
3	M3 copula: $C(u_1,u_2,u_3;\theta_1,\theta_2)=-\theta_1^{-1}\log\{1-(1-e^{-\theta_1})^{-1}(1-[1-(1-e^{-\theta_2})^{-1}(1-e^{-\theta_2 u_1})][(1-e^{-\theta_2 u_2})]^{(\theta_1/\theta_2)})(1-e^{-\theta_1 u_3})\};$ $\theta_2\ge\theta_1\in[0,\infty)$
4	M4 copula: $C(u_1,u_2,u_3;\theta_1,\theta_2)=[(u_1^{-\theta_2}+u_2^{-\theta_2}-1)^{(\theta_1/\theta_2)}+u_3^{-\theta_1}-1]^{(-1/\theta_1)};\theta_2\ge\theta_1\in[0,\infty)$
5	M5 copula: $C(u_1,u_2,u_3;\theta_1,\theta_2)=1-[\{(1-u_1)^{\theta_2}(1-(1-u_2)^{\theta_2})+(1-u_2)^{\theta_2}\}^{(\theta_1/\theta_2)}(1-(1-u_3)^{\theta_1})+(1-u_3)^{\theta_1}]^{(1/\theta_1)};$ $\theta_2\ge\theta_1\in[1,\infty)$
6	M6 copula: $C(u_1,u_2,u_3;\theta_1,\theta_2)=\exp\{-([(-\log u_1)^{\theta_2}+(-\log u_2)^{\theta_2}]^{(\theta_1/\theta_2)}+(-\log u_3)^{\theta_1})^{(1/\theta_1)}\};\theta_2\ge\theta_1\in[1,\infty)$ For M3, M4, M5 and M6 copulas, θ_1 and θ_2 are the dependence parameters

Goodness-of-fit tests were employed to select the best copula. We examined the null hypothesis $H_0: C\in C_0$ for a copula class C_0 against $H_1: C\notin C_0$ in the selection procedure. The tests compare the distance between the empirical distribution of copula, C_n and an estimation C_{θ_n} of C obtained under H_0 [Genest et al., 2009]. The empirical joint distribution was used as the reference for selecting the best copula. For instance, the empirical copula of bivariate (u_1,u_2) is defined by:

$$C_n(u_1,u_2)=\frac{1}{N}\sum_{i=1}^{N}1\ (F_1(x_{1i})\le u_1 \text{ and } F_2(x_{2j})\le u_2) \tag{5.3}$$

The goodness-of-fit test for the bivariate case can be analyzed using a distance measure:

$$\Omega = \sqrt{n}\left\{C_n\left(u_1,u_2\right) - C_{\theta_n}\left(u_1,u_2\right)\right\}; u_1,u_2 \in [0,1] \tag{5.4}$$

Using graphical plots and goodness of fit statistics [Genest et al. 2009], the best copulas for the multivariate models were selected.

5.4.3 Computation of the CDF-based Probabilistic Drought Index

The previous sections described construction of seven different cases, namely, three univariate marginals-based, three bivariate copula-based and a trivariate asymmetric Archimedean copula-based drought models. Indices such as SPI, SRI, MSDI, JDI are obtained by performing inverse Gaussian transformation to the CDF probabilities, however, there is a loss of information on uncertainty in drought classification. Additionally, adopting the CDF value directly as a drought indicator as shown in Figure 5.1 conveys the idea that the user is simply looking at $\mathrm{P}(X_1 \le x_1)$ or $\mathrm{P}(X_1 \le x_1, X_2 \le x_2)$ or $\mathrm{P}(X_1 \le x_1, X_2 \le x_2, X_3 \le x_3)$ for decision making on drought status. Therefore, in this study, the CDF values are retained for probabilistic drought classification using graphical models−specifically hidden Markov models (HMMs). The use of a CDF-based probabilistic drought index for watershed-scale drought studies has not been explored previously. Adopting a multivariable, multiscalar approach in the proposed framework can yield a drought index that performs that is useful for short and long term drought monitoring. Graphical model-based drought classification using HMMs allows drought/non-drought states to evolve based on the long term time series of indicator variables at that location. The main advantages of using HMM in drought classification

are: (i) the thresholds for drought classes are not arbitrarily decided, but determined by the data, (ii) probabilistic classification is achieved implying that uncertainty involved in the classification is available to the users, (iii) similarities in drought state evolution in the seven models could be explored.

5.4.3.1 Hidden Markov Models

Hidden Markov models (HMMs) are a class of graphical models. In a graphical model, variables are denoted by nodes and their dependencies are represented by edges (Jordan, 2004). When the graph structure comprises of hidden nodes with connections to observed nodes such that temporal dependencies exist between the hidden nodes, it is known as an HMM. In the context of the present study, the hidden nodes are the latent drought states (denoted by *q*), while the joint CDF of hydrologic variables constitute the observations (C, see Figure 5.1). HMMs have been used for drought applications by Mallya et al. [2013a] and Ramadas and Govindaraju [2014].

Detailed description of an HMM and its properties can be found in Rabiner [1989]. The hidden states are assumed to possess a first order Markov property, i.e. the probability of the system being in any future state is independent of past states given the present state. The hidden state at instant *t*, q_t, is therefore a discrete variable representing one of the K states.

The major characteristics of the HMM used in this study can be summarized as follows:

(i) Given the state of the system at time $t-1$, q_t is independent of previous states i.e. $P(q_t \mid q_{t-1}, q_{t-2}, \ldots, q_1) = P(q_t \mid q_{t-1})$. The state transition probability matrix can be defined as $A = \{a_{ij}\}$ where $a_{ij} = P(q_{t+1} = j \mid q_t = i), \quad 1 \le i, j \le K$. The following constraint applies for the transition probabilities: $\sum_{j=1}^{K} a_{ij} = 1; \quad 1 \le i \le K$.

(ii) Given the current state q_t, the observation at that instant C_t is conditionally independent of past observations, and the probability $P(C_t \mid q_t)$ is known as the emission distribution. The observations in this case are probabilities that fall in [0,1] range, and as a result, beta probability emission distributions are utilized. The matrix $B = \{\alpha_i, \beta_i\}$ represents the parameters of the beta distribution.

(iii) The initial state distribution, i.e., the probability that the drought state at the instant $t = 1$ $P(q_1)$ is given by $\pi = \{\pi_i\}$ s.t. $\pi_i = P(q_1 = i), \quad 1 \le i \le K$. Also, $\sum_{i=1}^{K} \pi_i = 1$ holds good for the initial probabilities.

Finally, the posterior probability of being in a particular drought state at time t, that aids in drought state classification is given by $P(q_t = i \mid C, B); 1 \le i \le K$. The detailed derivations of the posterior probabilities and parameter estimation procedure for beta-HMMs are available in Ramadas and Govindaraju [2014].

5.5 Results

5.5.1 Estimation of Joint Probabilities

The three input variables in the study were streamflows at the watershed outlet, and precipitation and soil moisture that were spatially lumped over the study watershed area. Anomalies of these variables denoted as X_1, X_2, X_3, respectively were used in the drought analysis. Extreme value, generalized extreme value (GEV), normal, and student's *t* distributions were fit to these inputs, and tested using two-sample Kolmogorov-Smirnov (K-S) hypothesis test. Table 5.2 lists the *p* values and K-S test statistic obtained in the three cases. The best fit distribution was chosen such that its calculated *p*-value is greater than the significance level of 0.05 and the maximum among all distributions' *p* values, and the corresponding K-S test statistic was the smallest. X_2 and X_3 are best fit by generalized extreme value and normal distributions, respectively, as indicated by the results—with large *p* and low test statistic values (shown in bold in Table 5.2). In the case of streamflow anomaly X_1, however, *p* values are less than the significance level 0.05, suggesting that X_1 does not belong to any of the tested distributions. Therefore, ranked probability series is used as its marginal distribution $F_1(x_1)$. The three univariate CDFs u_1, u_2, u_3 respectively, of X_1, X_2, X_3 are shown in Figure 5.2. These plots are useful for understanding different drought categories in univariate drought models. Graphical comparisons with the corresponding empirical CDFs assert the fit of the selected distributions.

Table 5.2 Two sample K-S hypothesis test results of fitting marginals to drought-related hydroclimatic variables where X_1 is streamflow anomaly, X_2 is precipitation anomaly, and X_3 is soil moisture anomaly. The best-fit distributions with highest *p* value (> 0.05) are indicated in bold

Variable	Distributions[#] and Parameters*	K-S Test Results	
		p value	Test statistic
X_1	EV: μ=60.17; σ=163.95	10^{-15}	0.21
	GEV: k=0.033; σ=76.04; μ=-46.08	0.006	0.09
	N: μ=10^{-15}; σ=105.64	10^{-9}	0.15
	T: ν=2.44	10^{-128}	0.61
X_2	EV: μ=24.69; σ=55.86	10^{-6}	0.13
	GEV: k=-0.048; σ=38.08; μ=-20.28	**0.56**	**0.04**
	N: μ=10^{-15}; σ= 46.19	10^{-5}	0.12
	T: ν=8.09	10^{-108}	0.56
X_3	EV: μ=25.98; σ=49.25	0.15	0.06
	GEV: k= -0.35; σ=55.56; μ=-17.12	0.75	0.03
	N: μ=-10^{-14}; σ= 53.26	**0.89**	**0.03**
	T: ν=25.94	10^{-82}	0.48

*Parameters: μ=location parameter; σ=scale parameter; k=shape parameter; ν=degrees of freedom

[#]Distributions: EV-extreme value, GEV-generalized extreme value, N-normal, T-student's t distribution

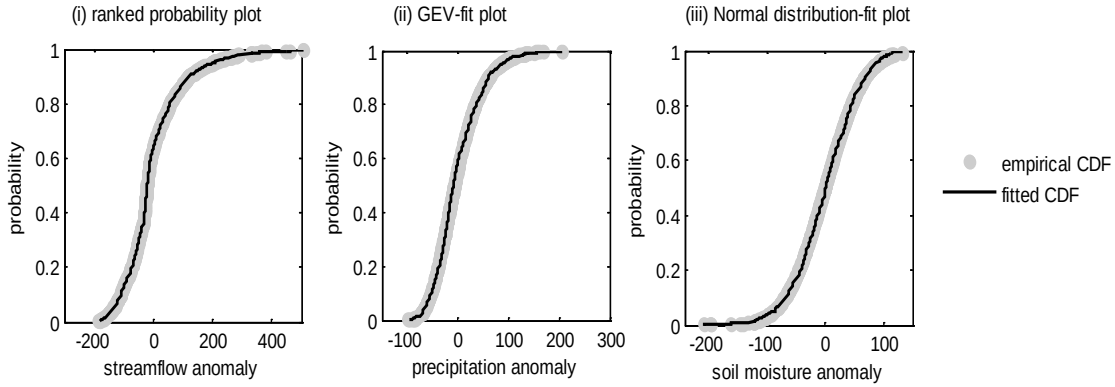

Figure 5.2 Comparison of CDF plots from empirical and best-fit distributions for (i) streamflow anomaly- using ranked probabilities, (ii) precipitation anomaly using GEV distribution and (iii) soil moisture anomaly using normal distribution

The best bivariate and trivariate copulas were selected using RMSE, and distance measure Ω-based Cramer-von-Mises (S_n) and Kolmogorov-Smirnov (T_n) statistics [Genest et al. 2009, Maity et al. 2013]. Table 5.3 lists the statistic values for each copula family, and the best copula selected has the smallest test statistic values. Gumbel copula has the best fit for bivariate copulas of (u_1, u_2) and (u_2, u_3), while the pair (u_3, u_1) is best fit by a Frank copula. Figure 5.3 shows the scatter plots of bivariate copula-generated data points with the observed points of all the three pairs. The selected copulas in each case have captured the observation space and also the variability, especially, in the extreme range. The plots also show the nature of correlation between the variable anomalies−correlation is maximum between the pair streamflow anomaly and soil moisture anomaly (Figure 5.3, plot iii).

Among the trivariate distributions tested in this study−Gaussian copula, student's *t*-copula, and asymmetric Archimedean or nested 3-copula families, the student's t-copula provides the best fit, based on goodness of fit statistics (provided in Table 5.3). The statistics RMSE, S_n and T_n are the lowest for this copula family. Similar to Figure 5.3, observed data points matched data points simulated using the best-fit copula, however, the plot is not included here for the sake of brevity. Figure 5.4 shows the plot of empirical and the best-fit copula CDFs. For ease of interpretation, data of different months of the year are shown by different symbols, and the selected student's *t*-copula fits the observed data well. Small discrepancies can be noted in the fit, as is seen for instance, in the monthly values for February, September, and November.

Table 5.3 Goodness-of-fit test results using Cramer-von-Mises statistic (S_n), Kolmogorov-Smirnov statistic (T_n), and root mean square error (RMSE) for the multivariate copula distributions used in the study. The best-fit cases are chosen based on low values of these statistics (shown in bold)

Copula*	$C(u_1,u_2)$			$C(u_2,u_3)$			$C(u_3,u_1)$		
	S_n	T_n	RMSE	S_n	T_n	RMSE	S_n	T_n	RMSE
CC	0.636	1.977	0.031	0.207	1.351	0.018	0.660	1.875	0.032
CF	0.109	1.126	0.013	0.078	0.925	0.011	**0.070**	**0.710**	**0.010**
CG	**0.071**	**0.839**	**0.010**	**0.061**	**0.793**	**0.010**	0.074	0.838	0.011
CT	0.105	0.982	0.013	0.066	0.870	0.010	0.080	0.821	0.011
CN	0.105	0.982	0.013	0.066	0.870	0.010	0.084	0.853	0.011

Copula*	$C(u_1,u_2,u_3)$		
	S_n	T_n	RMSE
M3	0.177	1.195	0.016
M4	1.200	2.843	0.043
M5	0.538	1.903	0.029
M6	0.204	1.470	0.018
CT	**0.144**	**1.074**	**0.015**
CN	0.145	1.070	0.015

* Note: CC-Clayton, CF-Frank, CG-Gumbel, CN-normal, CT-student's t copula
M3, M4, M5, M6 – nested 3-copula families.

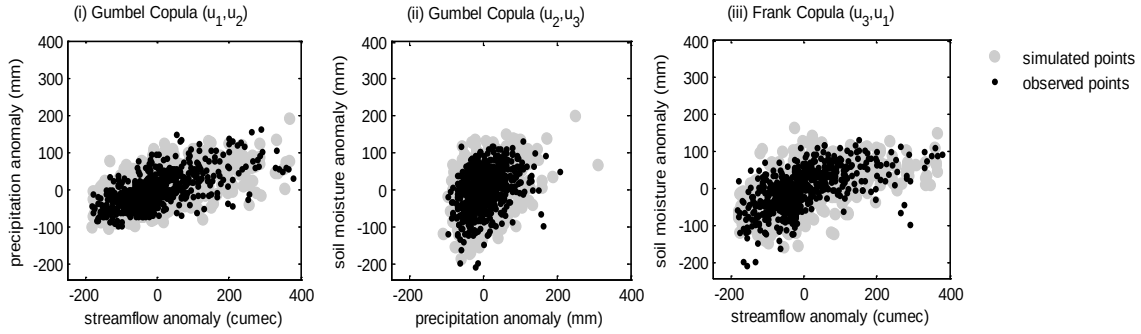

Figure 5.3 Comparison of available data points of variable anomalies (black dots) with simulated data points (gray circles) that were obtained using bivariate copulas: (i) and (ii) Gumbel copula for the pair streamflow anomaly and precipitation anomaly, and precipitation anomaly and soil moisture anomaly, respectively, and (iii) Frank copula for the pair soil moisture anomaly and streamflow anomaly

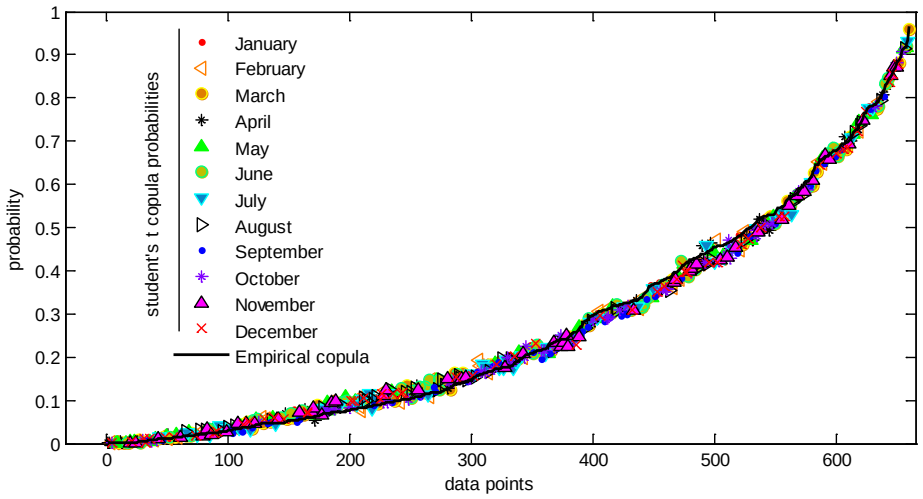

Figure 5.4 Cumulative distribution function (CDF) plots of the trivariate empirical copula (black line) and the selected student's *t* copula (different symbols are assigned for data of 12 months of the year). The selection was based on the goodness-of-fit statistics when multivariate student's t copula is compared with empirical CDF

5.5.2 CDF-based Probabilistic Drought Index

In the present study, the CDF probabilities from the seven drought models constitute the observations in 7 different HMMs, and five hidden drought/non-drought states were considered in each of these models. The state transitions were assumed to be smooth, allowing transitions to current state and neighboring states only. With the help of these assumptions and a sufficiently long time series, robust HMMs were constructed. The resulting hidden states from an HMM were characterized by the beta emission parameters α and β, and the initial state and transition state probabilities that were obtained after conducting several iterations (~100). The probabilities $P(q_t = i \mid O, B)$ indicate the evolving drought/non-drought state at each time step, and aid in assessing the uncertainty involved in the drought state classification.

The beta emission parameters and transition probabilities of the seven models were examined to understand the evolution of hidden states. The emission distribution parameters of the seven HMMs are listed in Table 5.4. The mean and variance corresponding to each beta emission distribution are also provided. The comparison allows us to comment on the performance of models, point out similarities in the evolution of states, and if a particular model could be selected as an overall drought indicator. Drought state 1 in the models is designated as a severe drought state as the corresponding CDF probabilities are the smallest (mean ≤ 0.1; see Table 5.4), indicating acute water deficits. State 2, in a similar fashion, is termed as a mild drought state because the probabilities represented by this beta distribution are small ($0.2 \leq \text{mean} \leq 0.5$), but indicate less severe deficit. These findings are substantiated by the CDF plots in Figure 5.1 for models 1 to 3. The variable anomaly values corresponding to these probabilities (defined by states 1 and 2) are negative. Using these plots, slightly larger probabilities ($0.5 \leq \text{mean} \leq 0.7$) falling in state 3 are attributed to normal conditions. The hidden states 4 and 5 corresponding to the larger CDF values ($0.7 \leq \text{mean} \leq 0.9$ and $0.9 \leq \text{mean} \leq 1$) are respectively, the mildly and severely wet states.

Comparison of model parameters of seven HMMs in Table 5.4 yields insights into nature of droughts represented by each category and each model. The emission model parameters suggest the shape and spread of a drought category that reflect the level of uncertainty in the class. There is greater agreement amongst models 4, 5, 6 and 7 that used multiple variables, in drought state classification (states 1 and 2). However, in the non-drought conditions (states 3, 4 and 5), the mean and variance values for models are different. In all the models, variance is maximum in the case of drought states 2 and 3

compared to the other states, as they cover a wider range of CDF probabilities that likely correspond to intermediate states suggested by standardized indices—such as moderate, mild and abnormal droughts, and normal to abnormally wet states, respectively. High variance models for drought state classification suggest large uncertainty. Out of the 7 models, model 3 representing univariate marginal of anomaly of soil moisture, model 6 corresponding to bivariate copula of streamflow and soil moisture anomalies, and model 7, the trivariate copula model of all the three variable anomalies show minimum uncertainty in drought state classification at 1-month time scale. The beta probability density functions (PDFs) of these 3 models are shown in Figure 5.5. The peaks associated with states 1 and 2 indicate that there is high probability of events in this category in the models. Similarly, a flat PDF, for instance, in the case of drought state 3, suggests large variance and increased uncertainty in the drought state classification. The shape and spread of emission distributions, reflect the propensity of droughts in each drought category.

The plots in Figure 2 explain how the CDF probabilities (C) falling under each of the 5 drought states (q) in models 1 to 3 can be translated into knowledge of the respective variable anomalies (X), and hence, the hydrologic conditions in the watershed. For instance, observations of model 3 are shown in Figure 5.2(iii). The five drought states can be better understood by juxtaposing CDF plots with the PDFs of beta emission distributions. For convenience, Figure 5.6 has CDF probabilities plotted against the input variable anomalies, representing models 3, 6 and 7, respectively. Using the plot in Figure 5.6(i), the means of emission distributions corresponding to five drought states of model 3 and corresponding streamflow anomaly thresholds can be examined. This is a

graphical representation of a CDF-based hydrologic drought index whose classification scheme was identified in the HMM formulation. Similarly, in Figure 5.6(ii), contours of bivariate copula CDF of streamflow anomaly and soil moisture anomaly input to model 6, aid in extracting variable anomaly values corresponding to each drought state. Mild drought state in this case can be seen to translate to streamflow and soil moisture anomaly values in the range -25 to -50 cumecs, and -75 to -100 mms respectively. Different drought states in model 7 can be inferred using Figure 5.6(iii), where differently colored points are shown in a three-dimensional plot of input variables. The mild and severe drought conditions suggested by model emission parameters correspond to the negative range of the three axes in this plot.

Table 5.5 lists the state transition probabilities for all the models in the drought classification scheme. The short term and long term drought monitoring capabilities of different models as well as the drought characteristics are reflected in the transition probabilities (Steinemann 2003). For instance, consider model 7, the first row has probabilities {0.72, 0.28, 0.00, 0.00, 0.00}, implying given that the current state is a severe drought, the most probable state (72%) at the next time step is severe drought itself, while 28% of the times there are likely transitions from this state to a less severe drought state. If we consider the mild drought state, the most probable category (61%) for transition at the next time step is remaining in the same state, and then, 13% and 27% of the times, respectively, transitioning to a severe drought and normal state. Persistence of the states (Steinemann, 2003) indicated by diagonal entries in the transition probability matrix is an important factor in drought planning. Persistence is quite low for the wet states in majority of the models (except in model 3), and for severe drought state in

models 2 and 4. Mildly wet state oscillates the least among the states in the models, with greater chances of transitioning to normal conditions. Mild drought, on the other hand, oscillates most among drought states, and in most cases, moves to normal condition in the next month (Table 5.5).

Table 5.4 HMM beta emission distribution parameters α and β for different dry/wet states in the one-month time scale drought classification models used in the study. Models 1,2,3, represent classification based on univariate marginals of anomalies of streamflows (X_1), precipitation (X_2), and soil moisture (X_3), respectively, and models 4, 5 and 6, correspond to bivariate copulas of pairs (X_1, X_2), (X_2, X_3) and (X_3, X_1), respectively. Model 7 used trivariate copula of (X_1, X_2, X_3)

Model	State 1				State 2				State 3			
	α	β	Mean	Variance	α	B	Mean	Variance	α	β	Mean	Variance
1	1	7	0.125	0.012	2	2	0.5	0.05	13	5	0.722	0.011
2	2	37	0.051	0.001	4	6	0.4	0.022	2	2	0.5	0.05
3	0.9	15	0.057	0.003	5	16	0.238	0.008	7	6	0.538	0.018
4	0.8	23	0.034	0.001	3	12	0.2	0.01	3	2	0.6	0.04
5	0.6	14	0.041	0.003	3	13	0.188	0.009	3	3	0.5	0.036
6	0.6	23	0.025	0.001	4	23	0.148	0.005	4	5	0.444	0.025
7	0.8	25	0.031	0.001	3	15	0.167	0.007	6	6	0.5	0.019

Model	State 4				State 5			
	α	β	Mean	Variance	α	β	Mean	Variance
1	44	3	0.936	0.001	485	2	0.996	0
2	56	6	0.903	0.001	5000	47	0.991	0
3	25	6	0.806	0.005	56	3	0.949	0.001
4	27	12	0.692	0.005	43	2	0.956	0.001
5	43	7	0.86	0.002	110	4	0.965	0
6	15	5	0.75	0.009	48	4	0.923	0.001
7	13	3	0.813	0.009	86	7	0.925	0.001

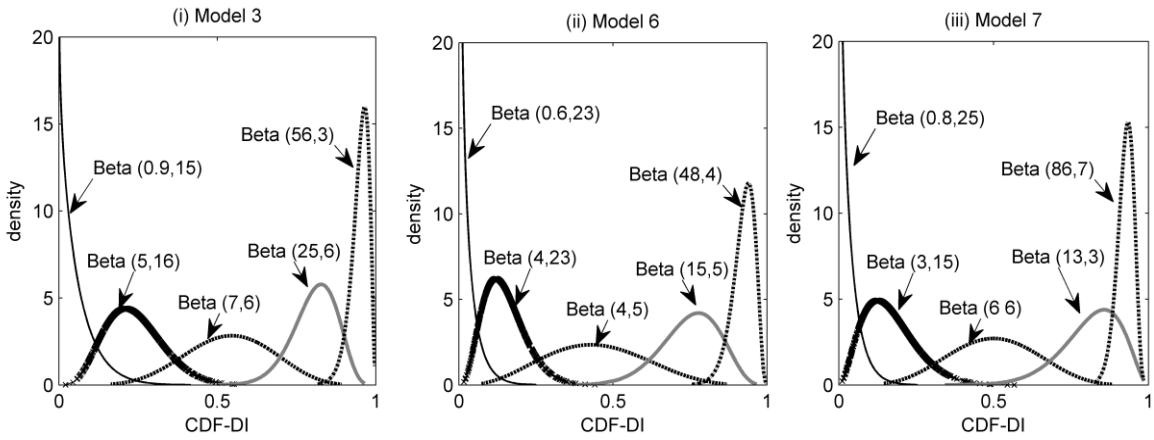

Figure 5.5 Sample PDF plots for the beta emission distributions corresponding to the five drought/non-drought states in (i) model 3, (ii) model 6 and (iii) model 7

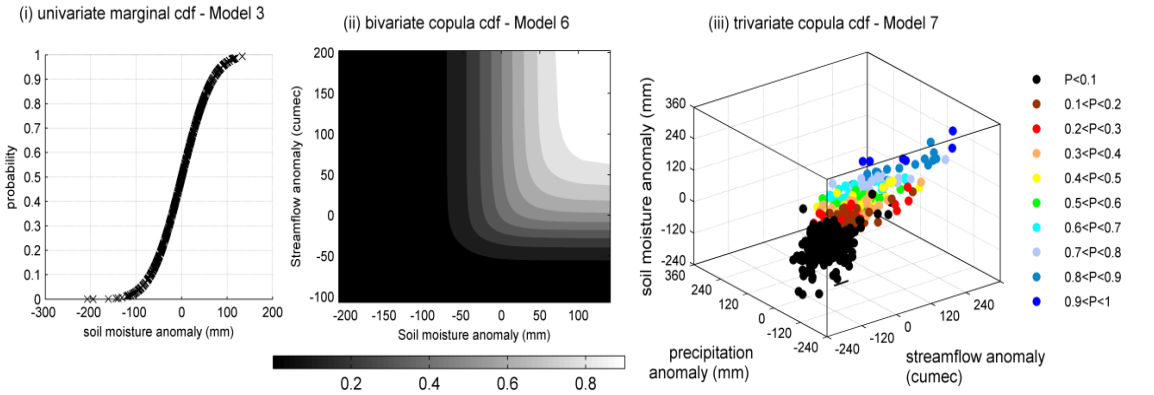

Figure 5.6 Sample CDF plots linking different probabilities in (i) model 3 (univariate), (ii) model 6 (bivariate) and (iii) model 7 (trivariate) to hydroclimatic anomalies

Table 5.5 HMM transition probabilities for different dry/wet states in the one-month time scale drought classification models used in the study. Models 1,2,3, represent classification based on univariate marginals of anomalies of streamflows (X_1), precipitation (X_2), and soil moisture (X_3), respectively, and models 4, 5 and 6, correspond to bivariate copulas of pairs (X_1, X_2), (X_2, X_3) and (X_3, X_1), respectively. Model 7 used trivariate copula of (X_1, X_2, X_3).

State	Model 1					Model 2					Model 3				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	0.60	0.40	0.00	0.00	0.00	0.33	0.67	0.00	0.00	0.00	0.74	0.26	0.00	0.00	0.00
2	0.11	0.80	0.09	0.00	0.00	0.29	0.12	0.59	0.00	0.00	0.11	0.68	0.21	0.00	0.00
3	0.00	0.32	0.44	0.24	0.00	0.00	0.13	0.74	0.13	0.00	0.00	0.15	0.70	0.15	0.00
4	0.00	0.00	0.72	0.20	0.08	0.00	0.00	0.85	0.14	0.01	0.00	0.00	0.26	0.61	0.13
5	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.35	0.65

State	Model 4					Model 5					Model 6				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	0.48	0.52	0.00	0.00	0.00	0.75	0.25	0.00	0.00	0.00	0.67	0.33	0.00	0.00	0.00
2	0.14	0.49	0.37	0.00	0.00	0.12	0.58	0.30	0.00	0.00	0.13	0.56	0.31	0.00	0.00
3	0.00	0.38	0.62	0.00	0.00	0.00	0.30	0.61	0.09	0.00	0.00	0.22	0.63	0.15	0.00
4	0.00	0.00	0.51	0.49	0.00	0.00	0.00	0.73	0.27	0.00	0.00	0.00	0.38	0.48	0.14
5	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.76	0.24

State	Model 7				
	1	2	3	4	5
1	0.72	0.28	0.00	0.00	0.00
2	0.13	0.60	0.27	0.00	0.00
3	0.00	0.41	0.46	0.13	0.00
4	0.00	0.00	0.75	0.25	0.00
5	0.00	0.00	0.00	1.00	0.00

5.5.3 Drought Classification

An HMM-based index not only extracts watershed-specific drought classes in this study, it has the added advantage of accounting for uncertainty in classification. The posterior probabilities of being in a particular state at any instant of time obtained from HMM reflect classification uncertainty. The results of drought classification at 1-month time scale for an example period 2001-2012 are provided in Figure 5.7. There are seven plots corresponding to each of the seven models. In each plot, the corresponding probabilities of falling in each drought state are shown using bars of different shades. The darkest shade corresponds to severe drought, whereas a white-colored bar represents a very wet state. The probabilities of being in each of the five states in a certain month indicate classification uncertainty, and the state that has the largest value is the most probable. At one-month time scale (Figure 5.7), only a few severe drought events have occurred in this region during 2001-2012 as indicated by all seven models−notable are those in the years 2001, 2007 and 2011-2012, that have been disastrous for the entire Midwest USA. Several mild droughts are reported by the models during this period. The smooth transitions imposed on the models are clearly visible in the drought evolution. Results suggest that models 2 and 3 yield the least number of drought instances. Model 1 recorded a large number of mild drought events during this period. Models 4, 5, 6, and 7 provide more realistic drought monitoring results−capable of both short term and long term drought management. Several drought events including those in 2001, 2007 and 2011-2012, were captured. Model 7−the trivariate case−is superfluous because model 5 gives similar results with just two inputs. In a similar fashion, models 1, 2, 6 and 7 are better suited for wet conditions.

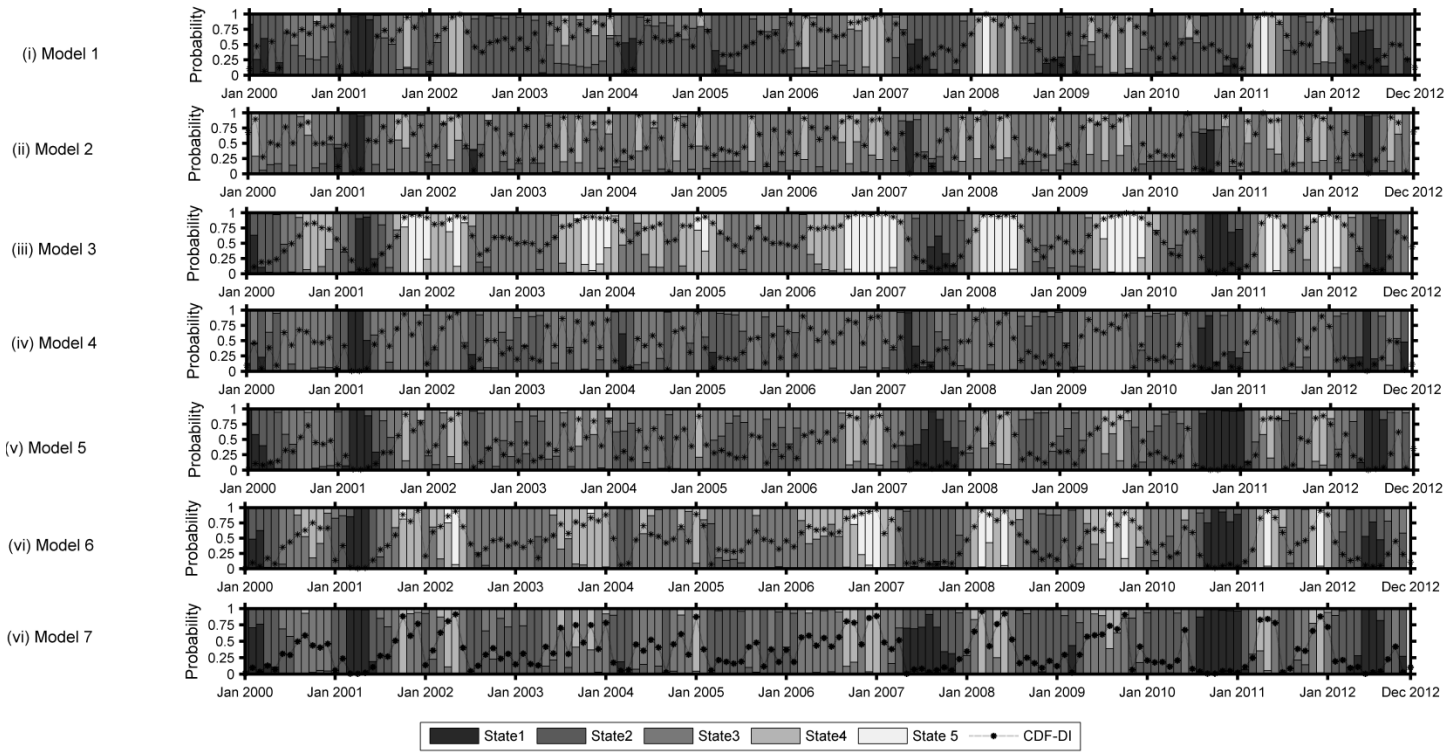

Figure 5.7 Probabilistic drought state classification by the proposed CDF-based index at one-month time scale in univariate and multivariate models 1 to 7 for the example period 2000-2012. Classification uncertainty is obtained since the probabilities of being in each of the four drought states are known rather than a single point estimate value of the drought index

5.5.3.1 Comparison of Models at Multiple Time Scales

For long term drought monitoring, drought indices are constructed at various time scales that are of interest to water managers. At 3- and 6-months scale, the responses to short term changes are less, unlike the 1-month time scale model. The emission parameters and transition state probabilities of the seven models at 3-month and 6-month time scales are provided in Tables B1-B4, included in Appendix B for brevity. The emission parameters in Tables B1 and B2 were compared with the one-month model (Table 5.4). There is reduction in variance of different drought classes as the time scale increases, implying less uncertainty in drought classification in these models. Low variance is a characteristic of the extreme states (severe drought and severely wet) in all models at all time scales, and these states are identified with high probabilities. Similarly, transition probabilities at 3- and 6-month time scales were examined. Persistence of states is high for all the states in the 3-month models 1, 3, 5 and 6 (Table B3), whereas, it is high in all seven 6-month drought models (Table B4), implying they are better indicators of long term drought conditions.

Transition probabilities for 3- and 6-month models in Tables B3 and B4 indicate the following trends in general: (i) persistence of states increase as the time scale increases, (ii) transitions from mild drought to normal conditions are observed with high probabilities, (iii) there are likely fewer transitions from mildly wet to severely wet conditions. Therefore, possible advantages of using 3- or 6-month time scale index for drought management are: (i) earlier identification of onset of drought, (ii) lower chances of false alarms, for instance if an abnormal drought indicator value is seen in a particular

month, (iii) drought responses may be triggered only when drought, that is, state 1 (severe) or 2 (mild) is encountered.

Probabilistic drought classification using 3-month and 6-month time scale models are shown in Figures 5.8 and 5.9. Overall, the posterior probabilities from the seven models for different drought classes are similar at 6-month time scale. In 3-month scale models, except for models 2 and 4, results from models are consistent. Upon closer examination, model 3, based on soil moisture anomaly is found to yield consistent drought monitoring results at 1-, 3-, and 6-month time scales. For all other models, there are differences between 1-month and the other two time scales. At 1- and 3-month time scales, the overall number of droughts captured is large. In Figure 5.9, results from 6-month scale models are shown, and few prominent long term mild and severe droughts captured by the models are in the years 2000, 2001, 2007, 2009, 2010-2011, 2012. Besides these, there are recorded droughts in 2002-2003, 2004, 2005-2006, 2008-2009, that are indicated by 1- and 3-month scale models alone. These observations are a key to understanding the utility of indices across time scales in the level of drought monitoring desired by users.

5.5.3.2 2012 Year Drought Outlook

The year 2012 was reported as a devastating drought year across whole of the Midwest USA, with severe consequences on the economy. A comparison across models and temporal scales for this particular drought is performed for understanding the 2012 drought evolution. If 1-month time scale models are considered, drought onset is observed as early as February 2012 across the 7 models (Figure 5.7). Model 3 indicates June as the first month of drought. In model 1, particularly, drought conditions persisted

in last 3-4 months of 2011, and drought in 2012 seems to have aggravated the deficit that was already in the system. The drought termination is observed very early on in models 2 and 4, before November.

Across 3-month scale models (Figure 5.8), the 2012 year droughts began mostly during March-May months, and early onsets are suggested by models 1, 2, 4 and 7. Only Model 5 suggests that conditions returned back to normal early, before November 2012. In 6-month scale drought models, earliest reported drought is in June 2012, and models except 2, 4 and 7 captured it one-two months later (Figure 5.9). The drought did not end before November 2012, according to these models.

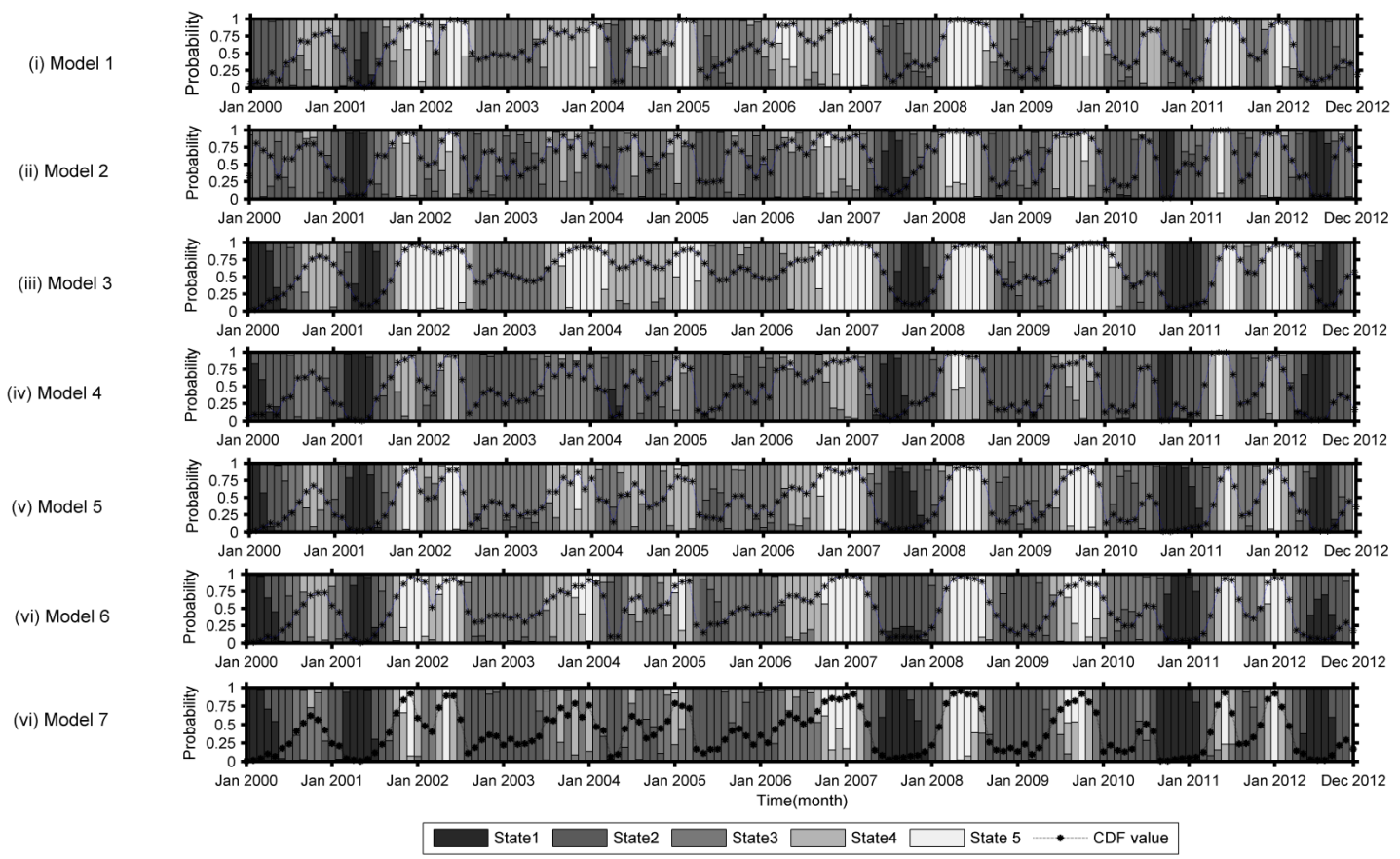

Figure 5.8 Probabilistic drought state classification by the proposed CDF-based index at 3-month time scale in univariate and multivariate models 1 to 7 for the example period 2000-2012. Classification uncertainty is obtained since the probabilities of being in each of the four drought states are known rather than a single point estimate value of the drought index

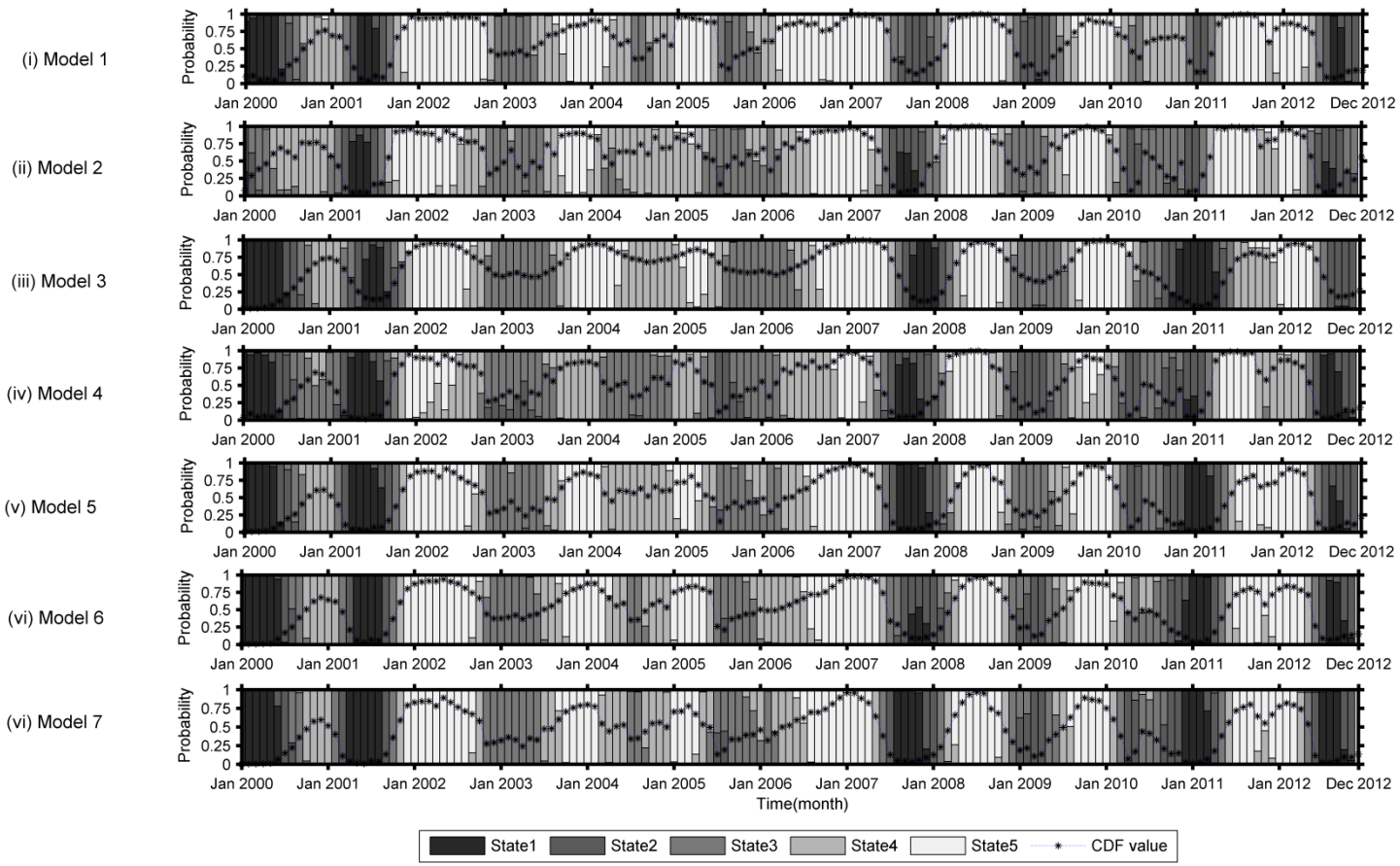

Figure 5.9. Probabilistic drought state classification by the proposed CDF-based index at 6-month time scale in univariate and multivariate models 1 to 7 for the example period 2000-2012. Classification uncertainty is obtained since the probabilities of being in each of the four drought states are known rather than a single point estimate value of the drought index

5.6 Summary and Conclusions

In this chapter, probabilistic drought indices that use univariate or multivariate CDF directly in an HMM framework were developed. The use of CDF estimate was suggested unlike the popular index formulations currently in use to allow for interpretation in terms of measured/modeled hydrological variables. Additionally, use of graphical models such as HMMs offers information on uncertainty in drought classification, that is, the probability of being in a given state at any time step is available. The analysis used lumped hydrological data over an Indiana watershed to develop univariate and multivariate drought models (total 7 in number) at three different timescales (1, 3, 6 months) for drought monitoring. The drought monitoring capabilities of the various models at different time scales were compared over an example 12 year time period.

The conclusions that can be drawn based on the case study are:

i. The indicators suggest probabilities that are easily translated to deficits in variables of interest. By means of contour plots and probability density function (PDF) plots, one can directly link drought categories to actual values of deficit in the meteorological, hydrological and agricultural systems. The CDF-based method allowed for inclusion of multiple hydrological variables without increasing the curse of dimensionality.

ii. The inclusion of multiple variables provided a multi-dimensional approach to drought characterization as indicated by the retrospective analysis using long term hydroclimatology that captured early drought onset, and persistent drought events in the region.

iii. Persistence in drought states increased as the time scale increased, implying 3- and 6-month models are better suited for long term drought monitoring.

iv. Models 4, 5 and 6, were parsimonious, with same drought detection capabilities as model 7, at all the three time scales. Bivariate joint models served as reasonably good overall drought indicators for this watershed. Among these, precipitation and streamflow are direct measurements.

v. Conforming to previous studies of drought indices at different time scales, short term droughts in the watershed are best captured by 1-month models, and fairly well by the 3-month models. Models at 6-months, however, picked only the prominent droughts that likely have serious widespread impacts on the watershed.

vi. Early onset of drought, as well as early withdrawal was suggested by one-month scale models, while the 6-month scale models reported the same drought months later. These differences are attributed to the cumulative nature of longer time scale models, and are useful to track deficits that are potential threats to long term water storage in the watershed.

vii. The evolving states in models are dependent on the long term hydro-climatology of the watershed, and are therefore watershed-specific.

Results of this case study have several implications on the understanding of various components of the hydrological cycle. The evolution of a watershed-scale drought from precipitation deficit, that leads to soil moisture deficit, and is ultimately reflected in streamflow deficit, explains the early onset of drought observed in the models based on precipitation, and persistence of droughts in streamflow drought models. Drought characterization evolved differently in the models considered in the case study. An integrated multivariate approach is beneficial for early drought monitoring, efficient mitigation and management, and this is achieved by determining the best subset of variables for index formulation through watershed-specific case studies.

CHAPTER 6. CONCLUSIONS

The focus of this research was to develop probabilistic models that have different applications in drought studies, namely, identifying drought triggers for hydrological droughts, predictor selection for drought models, developing a crop water stress-based agricultural drought index, and exploring roles of hydrologic variables for overall drought assessment at a watershed-scale. The applications used hydroclimatic variables such as streamflows, precipitation, soil moisture, temperature, runoff, evaporation, wind speed and sea level pressure at different locations in Indiana, USA.

6.1 Summary

Previous studies had suggested that there is scope for improving drought trigger-based information at a watershed scale [Palmer et al., 2002; Steinemann et al., 2005]. Though there are several retrospective drought characterization studies based on hydrologic data, there have been none on investigation of drought triggers. The first objective of this thesis used principal component analysis for dimensionality reduction, and copulas for joint probabilistic modeling, to extract triggers of hydrological droughts in two Indiana watersheds. The results of the study showed that drought triggers are watershed specific. Specific ranges of relevant hydro-climatic variables that are potential triggers to different categories of hydrologic droughts were extracted. Precipitation, soil

moisture, and runoff showed the greatest potential in resolving amongst different drought classes. These triggers are useful for forecasting expected value of streamflow deficit in watersheds at one-month lead time.

Predictor selection is another important aspect of hydrological modeling that has significant impacts on model performance and robustness. The second goal focused on determining the relevant predictors for parsimonious prediction of streamflows at any lead time. By using Gaussian graphical models based on conditional independence, a smaller subset of predictor variables was identified for prediction of streamflows at 1-, 2-, 3- and 4-month lead times. The resulting models performed as well as the models that used all the variables in the original set. The parsimonious streamflow prediction model at one-month lead time was then used for drought prediction at one-month time scale in the study watershed.

Agricultural drought studies in the past have utilized soil moisture as the primary drought indicator. However, the drought indices were not designed to account for crop responses to soil water deficits in the field. As the third goal, a new agricultural drought index was developed to account for crop water needs that are highly variable spatially and across the crop growing season duration, using crop water stress functions available from literature. Probabilistic classification of agricultural droughts was performed using graphical models (HMMs), where different hidden states represented different drought categories. The developed index suggested drought events that were in good agreement with results from popular indices such as SPI, PDSI and SPEI. Further, the propensity of severe and extreme category droughts across locations in Indiana was studied using the proposed index.

The thesis concludes with a case study to investigate the roles of hydrologic variables in overall drought assessment of a watershed. Bivariate and trivariate copulas were used to construct a joint CDF-based drought index as opposed to a multivariate analysis. Using different combinations of monthly precipitation, soil moisture and streamflows as drought indicators, and probabilistic classification using HMMs, indices were developed for different time scales (1, 3, and 6 months). The case study was useful in understanding how different hydrologic variables affect drought characterization and evolution. Inclusion of multiple variables captured the early onset of droughts as well as their persistence. Further, the models were watershed specific. Using a graphical model-based drought classification, the uncertainty involved in drought characterization was obtained.

Overall, probabilistic models were developed for applications in the field of drought trigger identification, drought prediction, monitoring, and classification. These drought models addressed some of the long standing questions in hydrologic studies such as dimensionality reduction, model parsimony, uncertainty estimation, and role of hydroclimatic variables in drought evolution. The confounding issues of availability of long record of data and model parameter space were tackled using PCA, copulas, and conditional independence. Graphical models proved to be a useful technique for model dimensionality reduction as well as drought classification with uncertainty estimation. In agricultural drought studies, use of crop water stress-based index was a new approach to capture drought events across space and time that vary because of cropping pattern and growth stage of crops, respectively. The results of the studies, in general, are watershed specific, and regional assessments are therefore not recommended.

6.2 Limitations of the Study

Few limitations of the drought research conducted in the thesis are as follows:

i. The hidden states in the HMM-based drought classification model that correspond to different drought categories are data driven, and therefore require interpretation. The mapping of drought classes using different models is by definition rather than from any underlying physics.

ii. Despite the efforts to reduce dimensionality of hydroclimatic predictors and use of simplifying assumptions in drought models, data limitations continue to affect model robustness in different applications. Parameter estimation is often dependent on initial estimates and requires multiple simulations with random starts, incurring large computational burden.

iii. Applications in the thesis primarily modeled the temporal evolution of droughts using probabilistic indices. However, spatio-temporal evolution of droughts still remains a challenging research problem.

iv. Most of the results were found to be watershed- or location-specific. Proper regionalization is not achieved in the modeling studies.

6.3 Future Work

Future research would call for techniques to improve probabilistic drought models to reduce classification uncertainty, and extending the range of applications for drought mitigation in watersheds. Examples of future research directions are listed as follows:

i. It is important to develop more robust drought models that can deal with sparse hydroclimatic data.

ii. While the proposed models are suitable for locations in Midwest USA, that is not the case for different locations around the world. When highly seasonal hydroclimatic variables are present, for instance in monsoon dominated regions, the proposed drought models for probabilistic drought classification need to be redesigned.

iii. To enhance the adoption and utility of the research by decision makers, web-based tools need to be developed for faster translation to application.

iv. Probabilistic analyses to address impacts of climate and land use change in watersheds remains an enduring challenge.

BIBLIOGRAPHY

BIBLIOGRAPHY

Alcamo, J., M. Flörke, and M. Marker. 2007. “Future Long-Term Changes in Global Water Resources Driven by Socio-Economic and Climatic Changes.” *Hydrological Sciences Journal* 52 (2): 247–275. doi:10.1623/hysj.52.2.247.

Anctil, F., N. Lauzon, and M. Filion. 2008. “Added Gains of Soil Moisture Content Observations for Streamflow Predictions Using Neural Networks.” *Journal of Hydrology* 359 (3-4): 225–234. doi:10.1016/j.jhydrol.2008.07.003.

Anmala, J., B. Zhang, and R. Govindaraju. 2000. “Comparison of ANNs and Empirical Approaches for Predicting Watershed Runoff.” *Journal of Water Resources Planning and Management* 126 (3): 156–166. doi:10.1061/(ASCE)0733-9496(2000)126:3(156).

Asefa, T., M. Kemblowski, M. McKee, and A. Khalil. 2006. “Multi-Time Scale Stream Flow Predictions: The Support Vector Machines Approach.” *Journal of Hydrology* 318 (1–4): 7–16. doi:http://dx.doi.org/10.1016/j.jhydrol.2005.06.001.

Aubert, D., C. Loumagne, and L. Oudin. 2003. “Sequential Assimilation of Soil Moisture and Streamflow Data in a Conceptual Rainfall–runoff Model.” *Journal of Hydrology* 280 (1–4): 145–161. doi:10.1016/S0022-1694(03)00229-4.

Bach, F. R., and M. I. Jordan. 2004. “Learning Graphical Models for Stationary Time Series.” *Signal Processing, IEEE Transactions on* 52 (8): 2189–2199. doi:10.1109/TSP.2004.831032.

Barnston, A. G. 1992. “Correspondence among the Correlation, RMSE, and Heidke Forecast Verification Measures; Refinement of the Heidke Score.” *Weather and Forecasting* 7 (4): 699-709.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss. 1970. “A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains.” *The Annals of Mathematical Statistics* 41 (1): 164–171.

Beguería, S., and S. M. Vicente Serrano. 2009. “SPEI Calculator.” Available at http://hdl.handle.net/10261/10002.

Bergman, K. H., P. Sabol, and D. Miskus. 1988. “Experimental Indices for Monitoring Global Drought Conditions.” In *Proceedings of the Thirteenth Annual Climate Diagnostics Workshop*, Cambridge, MA, U.S. Dept. of Commerce, 190–197.

Besaw, L. E., D M. Rizzo, P. R. Bierman, and W. R. Hackett. 2010. “Advances in Ungauged Streamflow Prediction Using Artificial Neural Networks.” *Journal of Hydrology* 386 (1–4): 27–37. doi:10.1016/j.jhydrol.2010.02.037.

Bishop, C. M., and M. E. Tipping. 2000. “Variational Relevance Vector Machines.” In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, C. Boutilier and M. Goldszmidt (Eds). Morgan Kaufmann Publishers Inc. Massachusetts, pp 46–53.

Bolten, J. D., W. T. Crow, X. Zhan, T. J. Jackson, and C. A. Reynolds. 2010. “Evaluating the Utility of Remotely Sensed Soil Moisture Retrievals for Operational Agricultural Drought Monitoring.” *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 3 (1): 57–66. doi:10.1109/JSTARS.2009 .2037163.

Bonev, B. 2010. “Feature Selection Based on Information Theory”. Ph.D. Thesis, University of Alicante, Alicante, Spain, 200pp.

Bonnin, G. M., D. Martin, B. Lin, T. Parzybok, M. Yekta, D. Riley. 2004. “Precipitation-frequency Atlas of the United States.” *NOAA Atlas 14* vol. 2. NOAA Natl. Weather Serv., Silver Spring, MD.

Bowden, G. J., G. C. Dandy, and H. R. Maier. 2005. “Input Determination for Neural Network Models in Water Resources Applications. Part 1—Background and Methodology.” *Journal of Hydrology* 301 (1-4): 75–92. doi:10.1016/j.jhydrol .2004.06.021.

Burget, L., P. Schwarz, M. Agarwal, P. Akyazi, Kai Feng, A. Ghoshal, O. Glembek, et al. 2010. “Multilingual Acoustic Modeling for Speech Recognition Based on Subspace Gaussian Mixture Models.” In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 4334–4337. doi:10.1109/ICASSP.2010.5495646.

Burn, D. H., J. M. Buttle, D. Caissie, G. MacCulloch, C. Spence, and K. Stahl. 2008. “The Processes, Patterns and Impacts of Low Flows across Canada.” *Canadian Water Resources Journal* 33 (2): 107–124. doi:10.4296/cwrj3302107.

Charusombat, U., and D. Niyogi. 2011. “A Hydroclimatological Assessment of Regional Drought Vulnerability: A Case Study of Indiana Droughts.” *Earth Interactions* 15 (26): 1–65. doi:10.1175/2011EI343.1.

Chen, Y., Q. Zhang, X. Chen, and P. Wang. 2012. “Multiscale Variability of Streamflow Changes in the Pearl River Basin, China.” *Stochastic Environmental Research and Risk Assessment* 26 (2): 235–246. doi:10.1007/s00477-011-0495-3.

Cover, T. M., and J. A. Thomas. 1991. *Elements of Information Theory*. Wiley, New York.

Crone, S. F., and N. Kourentzes. 2010. “Feature Selection for Time Series Prediction – A Combined Filter and Wrapper Approach for Neural Networks.” *Neurocomputing* 73 (10–12): 1923–1936. doi:http://dx.doi.org/10.1016/j.neucom.2010.01.017.

Crouse, M.S., R.D. Nowak, and R.G. Baraniuk. 1998. “Wavelet-Based Statistical Signal Processing Using Hidden Markov Models.” *Signal Processing, IEEE Transactions on* 46 (4): 886–902. doi:10.1109/78.668544.

Dai, A., K. E. Trenberth, and T. Qian. 2004. “A Global Dataset of Palmer Drought Severity Index for 1870–2002: Relationship with Soil Moisture and Effects of Surface Warming.” *Journal of Hydrometeorology* 5 (6): 1117–1130. doi:10.1175/JHM-386.1.

Davies, L., and U. Gather. 1993. “The Identification of Multiple Outliers.” *Journal of the American Statistical Association* 88 (423): 782–792.

Deheuvels, P. 1981. “A Non Parametric Test for Independence.” *Publ. de l’Inst. de Stat. de l’Univ. de Paris* 26, Inst. de Stat. Univ. de Paris, Paris, France, pp. 29–50.

Dempster, A. P. 1972. “Covariance Selection.” *Biometrics* 28 (1): 157-175.

Denmead, O. T., and R. H. Shaw. 1960. “The Effects of Soil Moisture Stress at Different Stages of Growth on the Development and Yield of Corn.” *Agronomy Journal* 52 (5): 272–274.

Dogan, E., S. Tripathi, D. A. Lyn, and R. S. Govindaraju. 2009. “From Flumes to Rivers: Can Sediment Transport in Natural Alluvial Channels Be Predicted from Observations at the Laboratory Scale?” *Water Resources Research* 45 (8). doi:10.1029/2008WR007637.

Doswell, C. A., R. Davies-Jones, and D. L. Keller. 1990. “On Summary Measures of Skill in Rare Event Forecasting Based on Contingency Tables.” *Weather and Forecasting* 5 (4): 576–585. doi:10.1175/1520-434(1990)005<0576:OSMOSI>2.0.CO;2.

Dracup, J. A., K. S. Lee, and E. G. Paulson. 1980. “On the Definition of Droughts.” *Water Resources Research* 16 (2): 297–302. doi:10.1029/WR016i002p00297.

Edwards, D. 2000. *Introduction to Graphical Modelling*. 2nd ed., Springer-Verlag, New York.

Edwards, D. C., and T. B. McKee. 1997. "Characteristics of 20th Century Drought in the United States at Multiple Time Scales." *Climatol. Rep., 97-2*, Dep. of Atmos. Sci., Colo. State Univ., Fort Collins.

Embrechts, P., F. Lindskog, and A. McNeil. 2003. "Modelling Dependence with Copulas and Applications to Risk Management." in *Handbook of Heavy Tailed Distributions in Finance*, Elsevier, New York, pp. 329–384.

Entin, J. K., A. Robock, K. Y. Vinnikov, S. E. Hollinger, S. Liu, and A. Namkhai. 2000. "Temporal and Spatial Scales of Observed Soil Moisture Variations in the Extratropics." *Journal of Geophysical Research: Atmospheres* 105 (D9): 11865–11877. doi:10.1029/2000JD900051.

Evans, R., D. K. Cassel, R. E. Sneed. 1996. *Soil, Water, and Crop Characteristics Important to Irrigation Scheduling*. AG 452-1, North Carolina Cooperative Extension Service, NC.

Fan, Y., and H. van den Dool. 2004. "Climate Prediction Center Global Monthly Soil Moisture Data Set at 0.5° Resolution for 1948 to Present." *Journal of Geophysical Research: Atmospheres* 109 (D10): D10102. doi:10.1029/2003JD004345.

Faul, A. C., and M. E. Tipping. 2001. "A Variational Approach to Robust Regression." In *Artificial Neural Networks — ICANN 2001*, edited by Georg Dorffner, Horst Bischof, and Kurt Hornik, 2130:95–102. Lecture Notes in Computer Science. Springer Berlin Heidelberg. Available at http://dx.doi.org/10.1007/3-540-44668-0_14.

Favre, A-C., S. El Adlouni, L. Perreault, N. Thiémonge, and B. Bobée. 2004. "Multivariate Hydrological Frequency Analysis Using Copulas." *Water Resources Research* 40 (1): W01101. doi:10.1029/2003WR002456.

Federal Emergency Management Agency. 1995. *National Mitigation Strategy; Partnerships for Building Safer Communities*, Mitigation Dir., Washington, D. C., 45 pp.

Fernando, T. M. K. G., H. R. Maier, and G. C. Dandy. 2009. "Selection of Input Variables for Data Driven Models: An Average Shifted Histogram Partial Mutual Information Estimator Approach." *Journal of Hydrology* 367 (3-4): 165–176. doi:10.1016/j.jhydrol.2008.10.019.

Fiori, M., P. Musé, and G. Sapiro. 2012. "Topology Constraints in Graphical Models." In *Advances in Neural Information Processing Systems* 25:800-808.

Galelli, S., and A. Castelletti. 2013. "Tree-Based Iterative Input Variable Selection for Hydrological Modeling." *Water Resources Research* 49 (7): 4295–4310. doi:10.1002/wrcr.20339.

Gao, C., M. Gemmer, X. Zeng, B. Liu, B. Su, and Y. Wen. 2010. "Projected Streamflow in the Huaihe River Basin (2010–2100) Using Artificial Neural Network." *Stochastic Environmental Research and Risk Assessment* 24 (5): 685–697. doi:10.1007/s00477-009-0355-6.

Genest, C., and A. Favre. 2007. "Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask." *Journal of Hydrologic Engineering* 12 (4): 347–368. doi:10.1061/(ASCE)1084-0699(2007)12:4(347).

Genest, C., B. Rémillard, and D. Beaudoin. 2009. "Goodness-of-Fit Tests for Copulas: A Review and a Power Study." *Insurance: Mathematics and Economics* 44 (2): 199–213. doi:10.1016/j.insmatheco.2007.10.005.

Genest, C., K. Ghoudi, and L.-P. Rivest. 1995. "A semiparametric estimation procedure of dependence parameters in multivariate families of distribution." *Biometrika* 82 (3): 543–552.

Georgakakos, K. P. 1986. “A generalized stochastic hydrometeorological model for flood and flash-flood forecasting: 1. Formulation.” *Water Resources Research* 22 (13): 2083–2095. doi: 10.1029/WR022i013p02083.

Ghosh, S., and P. P. Mujumdar. 2008. “Statistical Downscaling of {GCM} Simulations to Streamflow Using Relevance Vector Machine.” *Advances in Water Resources* 31 (1): 132–146. doi:http://dx.doi.org/10.1016/j.advwatres.2007.07.005.

Gibbons J. D., and S. Chakraborti. 2011. *Nonparametric Statistical Inference*, 5th ed., Chapman and Hall, Boca Raton, Florida.

Grimaldi, S., and F. Serinaldi. 2006. “Asymmetric Copula in Multivariate Flood Frequency Analysis.” *Advances in Water Resources* 29 (8): 1155–1167. doi:10.1016/j.advwatres.2005.09.005.

Han, W., Z. Yang, L. Di, and R. Mueller. 2012. “CropScape: A Web Service Based Application for Exploring and Disseminating {US} Conterminous Geospatial Cropland Data Products for Decision Support.” *Computers and Electronics in Agriculture* 84: 111–123. doi: http://dx.doi.org/10.1016/j.compag.2012.03.0 05.

Hao, Z., and A. AghaKouchak. 2013. “Multivariate Standardized Drought Index: A Parametric Multi-Index Model.” *Advances in Water Resources* 57 (July): 12–18. doi:10.1016/j.advwatres.2013.03.009.

Hao, Z., and A. AghaKouchak. 2014. “A Nonparametric Multivariate Multi-Index Drought Monitoring Framework.” *Journal of Hydrometeorology* 15 (1): 89–101. doi:10.1175/JHM-D-12-0160.1.

Hejazi, M. I., and X. Cai. 2009. “Input Variable Selection for Water Resources Systems Using a Modified Minimum Redundancy Maximum Relevance (mMRMR) Algorithm.” *Advances in Water Resources* 32 (4): 582–593. doi:http://dx.doi.org/10.1016/j.advwatres.2009.01.009.

Hocaoğlu, F. O., Ö. N. Gerek, and M. Kurban. 2010. “A Novel Wind Speed Modeling Approach Using Atmospheric Pressure Observations and Hidden Markov Models.” *Journal of Wind Engineering and Industrial Aerodynamics* 98 (8–9): 472–481. doi:http://dx.doi.org/10.1016/j.jweia.2010.02.003.

Holt, R. F., D. R. Timmons, W. B. Voorhees, and C. A. Van Doren. 1964. “Importance of Stored Soil Moisture to the Growth of Corn in the Dry to Moist Subhumid Climatic Zone.” *Agronomy Journal* 56 (1): 82–85.

Hoque, Y. M., S. Tripathi, M. M. Hantush, and R. S. Govindaraju. 2012. “Watershed Reliability, Resilience and Vulnerability Analysis under Uncertainty Using Water Quality Data.” *Journal of Environmental Management* 109: 101–112. doi:http://dx.doi.org/10.1016/j.jenvman.2012.05.010.

Hsiao, T. C. 1973. “Plant Responses to Water Stress.” *Annual Review of Plant Physiology* 24 (1): 519–570.

Hsu, C-N., H-J. Huang, and S. Dietrich. 2002. “The ANNIGMA-Wrapper Approach to Fast Feature Selection for Neural Nets.” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 32 (2): 207–212. doi:10.1109/3477.990877.

Huang, J., H. M. van den Dool, and K. P. Georgarakos. 1996. “Analysis of Model-Calculated Soil Moisture over the United States (1931–1993) and Applications to Long-Range Temperature Forecasts.” *Journal of Climate* 9 (6): 1350–1362. doi:10.1175/1520-0442(1996)009<1350:AOMCSM>2.0.CO;2.

Ihler, A. T., S. Kirshner, M. Ghil, A. W. Robertson, and P. Smyth. 2007. “Graphical Models for Statistical Inference and Data Assimilation.” *Physica D: Nonlinear Phenomena* 230 (1): 72–87.

Jensen, F. V., and Nielsen, T. D. 2007. *Bayesian Networks and Decision Graphs*. Springer-Verlag, New York.

Joe, H. 1997. *Multivariate Models and Dependence Concepts.* Chapman and Hall, London.

Jolliffe, I. T., and D. B. Stephenson. 2003. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley-Blackwell, Hoboken, New Jersey, 240 pp.

Jolliffe, I. T. 1986. *Principal Component Analysis.* Springer, New York.

Jordan, M. I. 2004. "Graphical Models." *Statistical Science* 19 (1): 140–155. doi:10.2307/4144379.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, et al. 1996. "The NCEP/NCAR 40-Year Reanalysis Project." *Bulletin of the American Meteorological Society* 77 (3): 437–471. doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Kao, S-C., and R. S. Govindaraju. 2008. "Trivariate Statistical Analysis of Extreme Rainfall Events via the Plackett Family of Copulas." *Water Resources Research* 44 (2): W02415. doi:10.1029/2007WR006261.

.Kao, S.-C., and R. S. Govindaraju. 2010a. "A Copula-Based Joint Deficit Index for Droughts." *J. Hydrol.*, 380(1–2): 121–134. doi:10.1016/ j.jhydrol.2009.10.029.

Kao, S.-C., and R. S. Govindaraju. 2010b. "Reply to Comment by T. P. Hutchinson on 'Trivariate statistical analysis of extreme rainfall events via Plackett family of copulas'." *Water Resour. Res*., 46: W04802. doi:10.1029/2009WR008774.

Karamouz, M., A. Ahmadi, and A. Moridi. 2009. "Probabilistic Reservoir Operation Using Bayesian Stochastic Model and Support Vector Machine." *Advances in Water Resources* 32 (11): 1588–1600. doi:http://dx.doi.org/10.1016/j.advwatres.2009.08.003.

Karamouz, M., K. Rasouli, and S. Nazif. 2009. “Development of a Hybrid Index for Drought Prediction: Case Study.” *Journal of Hydrologic Engineering* 14 (6): 617–627.

Karamouz, M., S. Torabi, and S. Araghinejad. 2004. “Analysis of Hydrologic and Agricultural Droughts in Central Part of Iran.” *Journal of Hydrologic Engineering* 9 (5): 402–414. doi:10.1061/(ASCE)1084-0699(2004)9:5(402).

Kerr, E. 2012. “Brutal Drought Depresses Agriculture, Thwarting US and Texas Economies.” *The Southwest Economy*, no. Q4: 10–13.

Keyantash, J. A., and J. A. Dracup. 2004. “An Aggregate Drought Index: Assessing Drought Severity Based on Fluctuations in the Hydrologic Cycle and Surface Water Storage.” *Water Resources Research* 40 (9): W09304. doi:10.1029/2003WR002610.

Keyantash, J., and J. A. Dracup. 2002. “The Quantification of Drought: An Evaluation of Drought Indices.” *Bulletin of the American Meteorological Society* 83 (8): 1167–1180. doi:10.1175/1520-0477(2002)083<1191:TQODAE>2.3.CO;2.

Khalil, A., M. N. Almasri, M. McKee, and J. J. Kaluarachchi. 2005. “Applicability of Statistical Learning Algorithms in Groundwater Quality Modeling.” *Water Resources Research* W05010. doi:10.1029/2004WR003608.

Kisi, O., and M. Cimen. 2011. “A Wavelet-Support Vector Machine Conjunction Model for Monthly Streamflow Forecasting.” *Journal of Hydrology* 399 (1–2): 132–140. doi:http://dx.doi.org/10.1016/j.jhydrol.2010.12.041.

Kojadinovic, I., and J. Yan. 2011. “A Goodness-of-Fit Test for Multivariate Multiparameter Copulas Based on Multiplier Central Limit Theorems.” *Statistics and Computing* 21 (1): 17–30. doi:10.1007/s11222-009-9142-y.

Koster, R. D., S. P. P. Mahanama, B. Livneh, D. P. Lettenmaier, and R. H. Reichle. 2010. "Skill in Streamflow Forecasts Derived from Large-Scale Estimates of Soil Moisture and Snow." *Nature Geosci* 3 (9): 613–616. doi:10.1038/ngeo944.

Koster, R. D., and M. J. Suarez. 2001. "Soil Moisture Memory in Climate Models." *Journal of Hydrometeorology* 2 (6): 558–570. doi:10.1175/1525-7541(2001)002<0558:SMMICM>2.0.CO;2.

Laio, F., A. Porporato, C. P. Fernandez-Illescas, and I. Rodriguez-Iturbe. 2001. "Plants in Water-Controlled Ecosystems: Active Role in Hydrologic Processes and Response to Water Stress: IV. Discussion of Real Cases." *Advances in Water Resources* 24 (7): 745–762.

Lakshmi, V., T. Piechota, U. Narayan, and C. Tang. 2004. "Soil Moisture as an Indicator of Weather Extremes." *Geophysical Research Letters* 31 (11). doi:10.1029/2004GL019930.

Lauritzen, S. L. 1996. *Graphical Models* (Vol. 17). Oxford University Press Inc., New York.

Leggetter, C. J., and P. C. Woodland. 1995. "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models." *Computer Speech & Language* 9 (2): 171–185. doi:http://dx.doi.org/10.1006/csla.1995.0010.

Leu, S.-S., and T. J. W. Adi. 2011. "Probabilistic prediction of tunnel geology using a Hybrid Neural-HMM." *Engineering Applications of Artificial Intelligence*, 24, 658–665.

Lins, H. F. 1985. "Interannual Streamflow Variability in the United States Based on Principal Components." *Water Resources Research* 21 (5): 691–701. doi:10.1029/WR021i005p00691.

Liu, W. T., and F. N. Kogan. 1996. “Monitoring Regional Drought Using the Vegetation Condition Index.” *International Journal of Remote Sensing* 17 (14): 2761–2782. doi:10.1080/01431169608949106.

Livneh, B., and D. P. Lettenmaier. 2012. “Multi-Criteria Parameter Estimation for the Unified Land Model.” *Hydrology and Earth System Sciences Discussions* 9 (4): 4417–4463. doi:10.5194/hessd-9-4417-2012.

Madadgar, S., and H. Moradkhani. 2013a. “Drought analysis under climate change using copula.” *J. Hydrol. Eng*. 746–759, doi:10.1061/ (ASCE)HE.1943–5584.0000532.

Madadgar, S., and H. Moradkhani. 2013b. “A Bayesian Framework for Probabilistic Seasonal Drought Forecasting.” *J. Hydrometeorology* 14: 1685–1705.

Madadgar, S., and H. Moradkhani. 2014. “Spatio-Temporal Drought Forecasting within Bayesian Networks.” *J. Hydrology* 512: 134–146.

Mahanama, S. P. P., R. D. Koster, R. H. Reichle, and L. Zubair. 2008. “The Role of Soil Moisture Initialization in Subseasonal and Seasonal Streamflow Prediction – A Case Study in Sri Lanka.” *Advances in Water Resources* 31 (10): 1333–1343. doi:10.1016/j.advwatres.2008.06.004.

Maier, H. R., A. Jain, G. C. Dandy, and K. P. Sudheer. 2010. “Methods Used for the Development of Neural Networks for the Prediction of Water Resource Variables in River Systems: Current Status and Future Directions.” *Environmental Modelling & Software* 25 (8): 891–909. doi:10.1016/j.envsoft.2010.02.003.

Maity, R., P. P. Bhagwat, and A. Bhatnagar. 2010. “Potential of Support Vector Regression for Prediction of Monthly Streamflow using Endogenous Property.” *Hydrological Processes* 24(7):917–923. doi: 10.1002/hyp.7535

Maity, R., A. Sharma, D. Nagesh Kumar, and K. Chanda. 2013. “Characterizing Drought Using the Reliability-Resilience-Vulnerability Concept.” *Journal of Hydrologic Engineering* 18 (7): 859–869. doi:10.1061/(ASCE)HE.1943-5584.0000639.

Maity, R., and S. S. Kashid. 2011. “Importance Analysis of Local and Global Climate Inputs for Basin-Scale Streamflow Prediction.” *Water Resources Research* 47 (11): W11504. doi:10.1029/2010WR009742.

Maity, R., and D. N. Kumar. 2008a. “Basin-Scale Stream-Flow Forecasting Using the Information of Large-Scale Atmospheric Circulation Phenomena.” *Hydrological Processes* 22 (5): 643–650. doi:10.1002/hyp.6630.

Maity, R., and D. N. Kumar.. 2008b. “Probabilistic Prediction of Hydroclimatic Variables with Nonparametric Quantification of Uncertainty.” *Journal of Geophysical Research: Atmospheres* 113 (D14): D14105. doi:10.1029/2008JD009856.

Maity, R., M. Ramadas, and R. S. Govindaraju. 2013. “Identification of Hydrologic Drought Triggers from Hydroclimatic Predictor Variables.” *Water Resources Research* 49 (7): 4476–4492.

Makkeasorn, A., N.B. Chang, and X. Zhou. 2008. “Short-Term Streamflow Forecasting with Global Climate Change Implications – A Comparative Study between Genetic Programming and Neural Network Models.” *Journal of Hydrology* 352 (3–4): 336–354. doi:10.1016/j.jhydrol.2008.01.023.

Malioutov, D. M., J. K. Johnson, and A. S. Willsky. 2006. “Walk-Sums and Belief Propagation in Gaussian Graphical Models.” *J. Mach. Learn. Res.* 7 (December): 2031–2064.

Mallya, G., S. Tripathi, S. Kirshner, and R. S. Govindaraju. 2013a. “Probabilistic Assessment of Drought Characteristics Using Hidden Markov Model.” *Journal of Hydrologic Engineering* 18 (7): 834–845. doi:10.1061/(ASCE)HE.1943-5584.0000699.

Mallya, G., L. Zhao, X. C. Song, D. Niyogi, and R. S. Govindaraju. 2013b. “2012 Midwest Drought in the United States.” *Journal of Hydrologic Engineering* 18 (7): 737–745. doi:10.1061/(ASCE)HE.1943-5584.0000786.

Manabe, S., and T. Delworth. 1990. “The Temporal Variability of Soil Wetness and Its Impact on Climate.” *Climatic Change* 16 (2): 185–192. doi:10.1007/BF00134656.

Margulis, S. A., D. McLaughlin, D. Entekhabi, and S. Dunne. 2002. “Land Data Assimilation and Estimation of Soil Moisture Using Measurements from the Southern Great Plains 1997 Field Experiment.” *Water Resources Research* 38 (12): 35–1–35–18. doi:10.1029/2001WR001114.

Maurer, E. P., D. P. Lettenmaier, and N. J. Mantua. 2004. “Variability and Potential Sources of Predictability of North American Runoff.” *Water Resources Research* 40 (9): W09306. doi:10.1029/2003WR002789.

May, R. J., H. R. Maier, G. C. Dandy, and T. M. K. G. Fernando. 2008. “Non-Linear Variable Selection for Artificial Neural Networks Using Partial Mutual Information.” *Environmental Modelling & Software* 23 (10-11): 1312–1326. doi:10.1016/j.envsoft.2008.03.007.

McHugh, M. J., and J. C. Rogers. 2001. “North Atlantic Oscillation Influence on Precipitation Variability around the Southeast African Convergence Zone.” *Journal of Climate* 14 (17): 3631–3642. doi:10.1175/1520-0442(2001)014<3631:NAOIOP>2.0.CO;2.

McKay, G. A., R. B. Godwin, and J. Maybank. 1989. “Drought and Hydrological Drought Research in Canada: An Evaluation of the State of the Art.” *Canadian Water Resources Journal* 14 (3): 71–84. doi:10.4296/cwrj1403071.

McKee, T. B., N. J. Doesken, and J. Kleist. 1993. “The Relationship of Drought Frequency and Duration to Time Scales.” In *Proceedings of the 8th Conference on Applied Climatology*, 17:179–83. held at Anaheim, California, LA, pp 179–183.

Meyer, S. J., K. G. Hubbard, and D. A. Wilhite. 1993. “A Crop-Specific Drought Index for Corn: II. Application in Drought Monitoring and Assessment.” *Agron. J.* 85 (2): 396–399.

Mishra, A. K., and V. P. Singh. 2010. “A Review of Drought Concepts.” *Journal of Hydrology* 391 (1–2): 202–216. doi:10.1016/j.jhydrol.2010.07.012.

Moghaddamnia, A., M. Ghafari Gousheh, J. Piri, S. Amin, and D. Han. 2009. “Evaporation Estimation Using Artificial Neural Networks and Adaptive Neuro-Fuzzy Inference System Techniques.” *Advances in Water Resources* 32 (1): 88–97. doi:http://dx.doi.org/10.1016/j.advwatres.2008.10.005.

Najjar, R. G. 1999. “The Water Balance of the Susquehanna River Basin and Its Response to Climate Change.” *Journal of Hydrology* 219 (1–2): 7–19. doi:10.1016/S0022-1694(99)00041-4.

Narasimhan, B., and R. Srinivasan. 2005. “Development and Evaluation of Soil Moisture Deficit Index (SMDI) and Evapotranspiration Deficit Index (ETDI) for Agricultural Drought Monitoring.” *Agricultural and Forest Meteorology* 133 (1–4): 69–88. doi:http://dx.doi.org/10.1016/j.agrformet.2005.07.012.

NDMC-UNL, 2012. The U.S. Drought Monitor is jointly produced by the National Drought Mitigation Center at the University of Nebraska-Lincoln, the United States Department of Agriculture, and the National Oceanic and Atmospheric Administration, available at: http://droughtmonitor.unl.edu/MapsandDataServices/MapService.aspx.

Nelsen, R. B. 2006. *An Introduction to Copulas.* Springer, New York.

Noori, R., A.R. Karbassi, A. Moghaddamnia, D. Han, M.H. Zokaei-Ashtiani, A. Farokhnia, and M. Ghafari Gousheh. 2011. “Assessment of Input Variables Determination on the SVM Model Performance Using PCA, Gamma Test, and Forward Selection Techniques for Monthly Stream Flow Prediction.” *Journal of Hydrology* 401 (3-4): 177–189. doi:10.1016/j.jhydrol.2011.02.021.

Oglesby, R. J., and D. J. Erickson. 1989. “Soil Moisture and the Persistence of North American Drought.” *Journal of Climate* 2 (11): 1362–1380. doi:10.1175/1520-0442(1989)002<1362:SMATPO>2.0.CO;2.

Palmer, W. C. 1965. *Meteorological Drought.* Weather Bureau Research Paper No. 45, U. S. Dept. of Commerce, Washington DC, 58 pp.

Palmer, W. C. 1968. “Keeping Track of Crop Moisture Conditions, Nationwide: The New Crop Moisture Index.” *Weatherwise* 21 (4): 156–161. doi:10.1080/00431672.1968.9932814.

Parthasarathy, B., K.Rupa Kumar, and A.A. Munot. 1993. “Homogeneous Indian Monsoon Rainfall: Variability and Prediction.” *Proceedings of the Indian Academy of Sciences - Earth and Planetary Sciences* 102 (1): 121–155. doi:10.1007/BF02839187.

Pearson, K. 1904. “On the Theory of Contingency and its Relation to Association and Normal Correlation.” *Draper’s Comp. Res. Mem. Biometric Ser. I.* Dulau and Co., London, U. K.

Peng, H., F. Long, and C. Ding. 2005. “Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy.” *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 27 (8): 1226–1238.

Phatak, A., B. C. Bates, and S. P. Charles. 2011. “Statistical Downscaling of Rainfall Data Using Sparse Variable Selection Methods.” *Environmental Modelling & Software* 26 (11): 1363–1371.

Porporato, A., F. Laio, L. Ridolfi, and I. Rodriguez-Iturbe. 2001. “Plants in Water-Controlled Ecosystems: Active Role in Hydrologic Processes and Response to Water Stress: III. Vegetation Water Stress.” *Advances in Water Resources* 24 (7): 725–744.

Potop, V., M. Možný, and J. Soukup. 2012. “Drought Evolution at Various Time Scales in the Lowland Regions and Their Impact on Vegetable Crops in the Czech Republic.” *Agricultural and Forest Meteorology* 156: 121–133. doi:http://dx.doi.org/10.1016/j.agrformet.2012.01.002.

Prasad, K., S. K. Dash, and U. C. Mohanty. 2010. “A Logistic Regression Approach for Monthly Rainfall Forecasts in Meteorological Subdivisions of India Based on DEMETER Retrospective Forecasts.” *International Journal of Climatology* 30 (10): 1577–1588. doi:10.1002/joc.2019.

Praskievicz, S., and H. Chang. 2009. “A Review of Hydrological Modelling of Basin-Scale Climate Change and Urban Development Impacts.” *Progress in Physical Geography* 33 (5): 650–671. doi:10.1177/0309133309348098.

Preisendorfer, R. W. 1988. *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, New York.

Rabiner, L. 1989. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.” *Proceedings of the IEEE* 77 (2): 257–286. doi:10.1109/5.18626.

Rajsekhar, D., V. P. Singh, and A. K. Mishra. 2014. “Multivariate Drought Index: An Information Theory Based Approach for Integrated Drought Assessment.” *Journal of Hydrology*. doi:10.1016/j.jhydrol.2014.11.031.

Ramadas, M., and R. S. Govindaraju. 2014. “Probabilistic Assessment of Agricultural Droughts Using Graphical Models.” *Journal of Hydrology* (In Press). doi:http://dx.doi.org/10.1016/j.jhydrol.2014.09.026.

Rhoads, F. M., and C. D. Yonts. 1991. “Irrigation Scheduling for Corn—Why and How.” In *National Corn Handbook*, NCH-20, USDA, Washington D. C.

Robertson, D. E., and Q. J. Wang. 2009. “Selecting Predictors for Seasonal Streamflow Predictions Using a Bayesian Joint Probability (BJP) Modelling Approach.” In *18th World IMACS/MODSIM Congress*. Anderssen, R.S., R.D. Braddock and L.T.H. Newham (eds), Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation, July 2009, Cairns, Australia, pp. 2377-2383. Available at http://metronu.ulb.ac.be/imacs/cairns/A6/robertson_de.pdf.

Rodriguez-Iturbe, I., P. D’Odorico, A. Porporato, and L. Ridolfi. 1999a. “On the Spatial and Temporal Links between Vegetation, Climate, and Soil Moisture.” *Water Resources Research* 35 (12): 3709–3722. doi:10.1029/1999WR900255.

Rodríguez-Iturbe, I., P. D’Odorico, A. Porporato, and L. Ridolfi. 1999b. “Tree-Grass Coexistence in Savannas: The Role of Spatial Dynamics and Climate Fluctuations.” *Geophysical Research Letters* 26 (2): 247–250. doi:10.1029/1998GL900296.

Ropelewski, C. F., and M. S. Halpert. 1996. “Quantifying Southern Oscillation-Precipitation Relationships.” *Journal of Climate* 9 (5): 1043–1059. doi:10.1175/1520-0442(1996)009<1043:QSOPR>2.0.CO;2.

Salas, J. D., C. Fu, and B. Rajagopalan. 2011. “Long-range Forecasting of Colorado Streamflows Based on Hydrologic, Atmospheric, and Oceanic Data.” *J. Hydrol. Eng.*, 16: 508–520, doi:10.1061/(ASCE)HE.1943–5584.0000343.

Salvadori, G., and C. De Michele. 2004. “Frequency Analysis via Copulas: Theoretical Aspects and Applications to Hydrological Events.” *Water Resources Research* 40 (12): W12511. doi:10.1029/2004WR003133.

Scholes, R. J., Walker, B. H. 1993. *An African Savanna: Synthesis of the Nylsvley Study*. Cambridge, UK: Cambridge Univ. Press.

Serinaldi, F., B. Bonaccorso, A. Cancelliere, and S. Grimaldi. 2009. "Probabilistic Characterization of Drought Properties through Copulas." *Physics and Chemistry of the Earth, Parts A/B/C* 34 (10-12): 596–605. doi:10.1016/j.pce.2008.09.004.

Shafer, B. A., and L. E. Dezman. 1982. "Development of a Surface Water Supply Index (SWSI) to Assess the Severity of Drought Conditions in Snowpack Runoff Areas." In *Proceedings of the 50th Annual Western Snow Conference*, 164–175. Reno, Nevada: Western Snow Conference. Available at sites/westernsnowconference.org/PDFs/1982Shafer.pdf.

Sharma, A. 2000. "Seasonal to Interannual Rainfall Probabilistic Forecasts for Improved Water Supply Management: Part 1—A Strategy for System Predictor Identification." *Journal of Hydrology* 239 (1): 232–239.

Sharma, A., K. C. Luk, I. Cordery, and U. Lall. 2000. "Seasonal to Interannual Rainfall Probabilistic Forecasts for Improved Water Supply Management: Part 2—Predictor Identification of Quarterly Rainfall Using Ocean-Atmosphere Information." *Journal of Hydrology* 239 (1): 240–248.

Sheffield, J., G. Goteti, F. Wen, and E. F. Wood. 2004. "A Simulated Soil Moisture Based Drought Analysis for the United States." *Journal of Geophysical Research: Atmospheres* 109 (D24). doi:10.1029/2004JD005182.

Sheffield, J., and E. F. Wood. 2008. "Global Trends and Variability in Soil Moisture and Drought Characteristics, 1950–2000, from Observation-Driven Simulations of the Terrestrial Hydrologic Cycle." *Journal of Climate* 21 (3): 432–458. doi:10.1175/2007JCLI1822.1.

Shiau, J. T. 2006. “Fitting Drought Duration and Severity with Two-Dimensional Copulas.” *Water Resources Management* 20 (5): 795–815. doi:10.1007/s11269-005-9008-9.

Shiau, J-T., Song, F., and S. Nadarajah. 2007. “Assessment of Hydrological Droughts for the Yellow River, China, Using Copulas.” *Hydrological Processes* 21 (16): 2157–2163. doi:10.1002/hyp.6400.

Shukla, S., and A. W. Wood. 2008. “Use of a Standardized Runoff Index for Characterizing Hydrologic Drought.” *Geophysical Research Letters* 35 (2). doi:10.1029/2007GL032487.

Sklar, A. 1959. Fonction de répartition à n dimensions et leurs marges. Publications de Institut de Statistique Université de Paris 8, 229–231.

Srinivasan, R., and J. G. Arnold. 1994. “Integration of a Basin-Scale Water Quality Model with GIS.” *JAWRA Journal of the American Water Resources Association* 30 (3): 453–462. doi:10.1111/j.1752-1688.1994.tb03304.x.

Steinemann, A. 2003. “Drought Indicators and Triggers: A Stochastic Approach to Evaluation.” *JAWRA Journal of the American Water Resources Association* 39 (5): 1217–1233. doi:10.1111/j.1752-1688.2003.tb03704.x.

Steinemann, A. C., M. J. Hayes, and L. F. N. Cavalcanti. 2005. “Drought Indicators and Triggers.” In D.A. Wilhite (Ed.), *Drought and Water Crises: Science, Technology, and Management Issues*, 71–92. Boca Raton, FL: CRC Press.

Sun, W., H. Zhang, A. Palazoglu, A. Singh, W. Zhang, and S. Liu. 2013. “Prediction of 24-Hour-Average PM2.5 Concentrations Using a Hidden Markov Model with Different Emission Distributions in Northern California.” *Science of The Total Environment* 443 : 93–103. doi:http://dx.doi.org/10.1016/j.scitotenv.2012.10.070.

Svoboda, M., D. LeComte, M. Hayes, R. Heim, K. Gleason, J. Angel, B. Rippey, et al. 2002. “The Drought Monitor.” *Bulletin of the American Meteorological Society* 83 (8): 1181–1190. doi:10.1175/1520-0477(2002)083<1181:TDM>2.3.CO;2.

Tang, C., and T. C. Piechota. 2009. “Spatial and Temporal Soil Moisture and Drought Variability in the Upper Colorado River Basin.” *Journal of Hydrology* 379 (1–2): 122–135. doi:http://dx.doi.org/10.1016/j.jhydrol.2009.09.052.

Templeton, G. F. 2011. “A Two-Step Approach for Transforming Continuous Variables to Normal: Implications and Recommendations for IS Research.” *Communications of the Association for Information Systems* 28(4). Available at: http://aisel.aisnet.org/cais/ vol28/iss1/4.

Thornthwaite, C. W. 1948. “An Approach toward a Rational Classification of Climate.” *Geographical Review* 38 (1): 55–94.

Tian, Y., M. J. Booij, and Y-P. Xu. 2014. “Uncertainty in High and Low Flows due to Model Structure and Parameter Errors.” *Stochastic Environmental Research and Risk Assessment* 28 (2): 319–332. doi:10.1007/s00477-013-0751-9.

Tolk, J. A. 2003. “Soils, Permanent Wilting Points.” Encyclopedia of Water Science, doi:10.1081/E-EWS 120010337, Marcel Dekker, Inc.

Traveria, M., A. Escribano, and P. Palomo. 2010. “Statistical Wind Forecast for Reus Airport.” *Meteorological Applications* 17 (4): 485–495. doi:10.1002/met.192.

Trenberth, K. E. 1999. “Conceptual Framework for Changes of Extremes of the Hydrological Cycle with Climate Change.” *Climatic Change* 42 (1): 327–339. doi:10.1023/A:1005488920935.

Tripathi, S., and R. Govindaraju. 2011. “Appraisal of Statistical Predictability under Uncertain Inputs: SST to Rainfall.” *Journal of Hydrologic Engineering* 16 (12): 970–983. doi:10.1061/(ASCE)HE.1943-5584.0000278.

Tripathi, S., and R. S. Govindaraju. 2008. “Engaging Uncertainty in Hydrologic Data Sets Using Principal Component Analysis: BaNPCA Algorithm.” *Water Resources Research* 44 (10): W10409. doi:10.1029/2007WR006692.

Tripathi, S., and R. S. Govindaraju. 2007. “On Selection of Kernel Parameters in Relevance Vector Machines for Hydrologic Applications.” *Stochastic Environmental Research and Risk Assessment* 21 (6): 747–764. doi:10.1007/s00477-006-0087-9.

Tripathi, S., V. V. Srinivas, and R. S. Nanjundiah. 2006. “Downscaling of Precipitation for Climate Change Scenarios: A Support Vector Machine Approach.” *Journal of Hydrology* 330 (3–4): 621–640. doi:http://dx.doi.org/10.1016/j.jhydrol.2006.04.030.

Vereecken, H., J. A. Huisman, H. Bogena, J. Vanderborght, J. A. Vrugt, and J. W. Hopmans. 2008. “On the Value of Soil Moisture Measurements in Vadose Zone Hydrology: A Review.” *Water Resources Research* 44 (4). doi:10.1029/2008WR006829.

Vicente-Serrano, S. M., S. Beguería, and J. I. López-Moreno. 2010. “A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index.” *Journal of Climate* 23 (7): 1696–1718. doi:10.1175/2009JCLI2909.1.

Vicente-Serrano, S. M., S. Beguería, J. Lorenzo-Lacruz, J. J. Camarero, J. I. López-Moreno, C. Azorin-Molina, J. Revuelto, E. Morán-Tejeda, and A. Sanchez-Lorenzo. 2012. “Performance of Drought Indices for Ecological, Agricultural, and Hydrological Applications.” *Earth Interactions* 16 (10): 1–27. doi:10.1175/2012EI000434.1.

Wang, W-C., K-W. Chau, C-T. Cheng, and L. Qiu. 2009. “A Comparison of Performance of Several Artificial Intelligence Methods for Forecasting Monthly Discharge Time Series.” *Journal of Hydrology* 374 (3–4): 294–306.

Ward, M. N. 1992. “Provisionally Corrected Surface Wind Data, Worldwide Ocean-Atmosphere Surface Fields, and Sahelian Rainfall Variability.” *Journal of Climate* (United States) 5(5):454-475.

Weaver, J. E. and Bruner, W. E. 1927. *Root Development of Vegetable Crops*. First edition, McGraw-Hill Book Company, Inc., New York, available at: http://www.soilandhealth.org/01aglibrary/010137veg.roots/010137toc.html .

Weaver J. E. 1926. *Root Development of Field Crops*. McGraw-Hill Book Company Inc., New York, available at: http://www.soilandhealth.org/01aglibrary/010139fieldcroproots / 010139toc.html.

Wells, N., S. Goddard, and M. J. Hayes. 2004. “A Self-Calibrating Palmer Drought Severity Index.” *Journal of Climate* 17 (12): 2335–2351.

Western, A. W., R. B. Grayson, and T. R. Green. 1999. “The Tarrawarra Project: High Resolution Spatial Measurement, Modelling and Analysis of Soil Moisture and Hydrological Response.” *Hydrological Processes* 13 (5): 633–652.

Whelan, N. 2004. “Sampling from Archimedean Copulas.” *Quantitative Finance* 4 (3): 339–352. doi:10.1088/1469-7688/4/3/009.

Whittaker, Joe. 2009. *Graphical Models in Applied Multivariate Statistics*. Wiley Publishing. New York.

Wilks, D. S. 2006. *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press/Elsevier, New York, 627 pp.

Willsky, A. S. 2002. “Multiresolution Markov Models for Signal and Image Processing.” *Proceedings of the IEEE* 90 (8): 1396–1458. doi:10.1109/JPROC.2002.800717.

Wong, G., M. Lambert, M. Leonard, and A. Metcalfe. 2009. “Drought Analysis Using Trivariate Copulas Conditional on Climatic States.” *Journal of Hydrologic Engineering* 15 (2): 129–41. doi:10.1061/(ASCE)HE.1943-5584.0000169.

Wu, C. L., K. W. Chau, and Y. S. Li. 2009. “Predicting Monthly Streamflow Using Data-Driven Models Coupled with Data-Preprocessing Techniques.” *Water Resources Research* 45 (8): W08432. doi:10.1029/2007WR006737.

Wu, J., B. He, A. Lü, L. Zhou, M. Liu, and L. Zhao. 2011. “Quantitative Assessment and Spatial Characteristics Analysis of Agricultural Drought Vulnerability in China.” *Natural Hazards* 56 (3): 785–801. doi:10.1007/s11069-010-9591-9.

Xu, C.-Y., and V. P. Singh. 2004. “Review on Regional Water Resources Assessment Models under Stationary and Changing Climate.” *Water Resources Management* 18 (6): 591–612. doi:10.1007/s11269-004-9130-0.

Yau, C., O. Papaspiliopoulos, G. O. Roberts, and C. Holmes. 2011. “Bayesian Non-Parametric Hidden Markov Models with Applications in Genomics.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (1): 37–57.

Yu, H., Z. Choo, W. I. T. Uy, J. Dauwels, and P. Jonathan. 2012. “Modeling Extreme Events in Spatial Domain by Copula Graphical Models.” In *Information Fusion (FUSION), 2012 15th International Conference on*, 1761–1768.

Yu, M., Q. Li, M. J. Hayes, M. D. Svoboda, and R. R. Heim. 2013. “Are Droughts Becoming More Frequent or Severe in China Based on the Standardized Precipitation Evapotranspiration Index: 1951–2010?” *International Journal of Climatology*. doi:10.1002/joc.3701.

Zhang, H., W. Zhang, A. Palazoglu, and W. Sun. 2012. “Prediction of Ozone Levels Using a Hidden Markov Model (HMM) with Gamma Distribution.” *Atmospheric Environment* 62: 64–73. doi:http://dx.doi.org/10.1016/j.atmosenv.2012.08.008.

Zhang, L., and V. Singh. 2006. “Bivariate Flood Frequency Analysis Using the Copula Method.” *Journal of Hydrologic Engineering* 11 (2): 150–64. doi:10.1061/(ASCE)1084-0699(2006)11:2(150).

APPENDICES

Appendix A

Parameter Estimation using EM Algorithm

Given that observations in the time series and underlying sequence of states are represented as $O = (o_1, o_2, \ldots, o_T)$ and $q = (q_1, q_2, \ldots q_T)$, respectively, the log likelihood function to be maximized becomes:

$$Q(\mathrm{M}, \mathrm{M}') = \sum_q P(O, q \mid \mathrm{M}') \log P(O, q \mid \mathrm{M}) \tag{A.1}$$

where M represents the new set of model parameters and M' the previous/initial set of values. If we define the probability $P(O, q \mid \mathrm{M})$ as follows:

$$P(O, q \mid \mathrm{M}) = \pi_{q_1} \prod_{t=2}^{T} a_{q_{t-1} q_t} b_{q_t}(o_t) \tag{A.2}$$

where π, a and b denote initial state, transition and emission probabilities respectively.

Then, Q may be written as:

$$\begin{aligned} Q(\mathrm{M}, \mathrm{M}') &= \sum_q P(O, q \mid \mathrm{M}') \log \pi_{q_1} + \sum_q \sum_{t=2}^{T} \log a_{q_{t-1} q_t} P(O, q \mid \mathrm{M}') \\ &\quad + \sum_q \sum_{t=1}^{T} \log b_{q_t}(o_t) P(O, q \mid \mathrm{M}') \\ &= Q(I) + Q(II) + Q(III) \end{aligned} \tag{A.3}$$

Estimation of initial state probabilities π_i

The three parts of Equation (A.3) can be used to maximize the Q function; each term is optimized independently to obtain the new set of parameters.

For the first term $Q(I)$ in Equation (A.3), its maximization subject to constraint $\sum_i \pi_i = 1$ to obtain estimation formula for π_i follows:

$$\frac{\partial}{\partial \pi_i}\left(\sum_{i=1}^{K} \log \pi_i \; p(O, q_1 = i \mid \mathrm{M}') + \lambda_1 (\sum_{i=1}^{K} \pi_i - 1) \right) = 0$$
$$\Rightarrow \pi_i = \frac{P(O, q_1 = i \mid \mathrm{M}')}{P(O \mid \mathrm{M}')} = \frac{\alpha^*{}_1'(i)\beta^*{}_1'(i)}{\sum_{i=1}^{K} \alpha^*{}_1'(i)\beta^*{}_1'(i)} \tag{A.4}$$

where λ_1 is the Lagrange multiplier, and functions α^*, β^* are as defined in the forward-backward algorithm of Rabiner [1989]. Note that these functions are different from the beta emission parameters.

Estimation of transition state probabilities a_{ij}

Similar to the previous exercise, maximization of $Q(II)$ subject to constraint $\sum_{j=1}^{K} a_{ij} = 1$ is performed as follows:

$$\frac{\partial}{\partial a_{ij}}\left(\sum_{q} \sum_{t=2}^{T} \log a_{q_{t-1} q_t} P(O, q \mid \mathrm{M}') \right) + \lambda_2 (\sum_{j=1}^{K} a_{ij} - 1 = 0)$$
$$= \frac{\partial}{\partial a_{ij}}\left(\sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{t=2}^{T} \log a_{ij} P(O, q_{t-1=i}, q_{t=j} \mid \mathrm{M}') \right) + \lambda_2 (\sum_{j=1}^{K} a_{ij} - 1 = 0) \tag{A.5}$$
$$\Rightarrow a_{ij} = \frac{\sum_{t=2}^{T} P(O, q_{t-1=i}, q_{t=j} \mid \mathrm{M}')}{\sum_{t=2}^{T} P(O, q_{t-1=i} \mid \mathrm{M}')} = \frac{\sum_{t=2}^{T} \alpha^*{}_{t-1}'(i) a_{ij}' \beta^*{}_t'(j) b_j'(o_t)}{\sum_{j=1}^{K} \sum_{t=2}^{T} \alpha^*{}_{t-1}'(i) a_{ij}' \beta^*{}_t'(j) b_j'(o_t)}$$

Estimation of beta emission distribution parameters

Maximizing $Q(III)$ does not involve Lagrange multipliers as there are no constraints for the beta emission distribution parameters α_j, β_j.

$$\frac{\partial}{\partial \alpha_i}\left(\sum_q\sum_{t=1}^{T}\log b_{q_t}(o_t)P(O,q\,|\,\mathrm{M}')\right)=\left(\sum_q\sum_{t=1}^{T}P(O,q\,|\,\mathrm{M}')\frac{\partial\left(\log b_{q_t}(o_t)\right)}{\partial \alpha_i}\right)=0$$
$$\frac{\partial}{\partial \beta_i}\left(\sum_q\sum_{t=1}^{T}\log b_{q_t}(o_t)P(O,q\,|\,\mathrm{M}')\right)=\left(\sum_q\sum_{t=1}^{T}P(O,q\,|\,\mathrm{M}')\frac{\partial\left(\log b_{q_t}(o_t)\right)}{\partial \beta_i}\right)=0 \qquad \text{(A.6)}$$

Emission density for beta distribution is $b_j(O_t)=betapdf(o_t,\alpha_j,\beta_j)=\frac{{o_t}^{\alpha_j-1}(1-o_t)^{\beta_j-1}}{\mathrm{B}(\alpha_j,\beta_j)}$.

Derivatives in Equation (A.6) can be expanded as:

$$\frac{\partial\left(\log b_{q_t}(o_t)\right)}{\partial \alpha_j}=\frac{1}{b_{q_t}(o_t)}\frac{\partial\left(b_{q_t}(o_t)\right)}{\partial \alpha_j}=\frac{1}{b_{q_t}(o_t)}\cdot\frac{\partial}{\partial \alpha_j}\left(\frac{{o_t}^{\alpha_j-1}(1-o_t)^{\beta_j-1}}{\mathrm{B}(\alpha_j,\beta_j)}\right)$$
$$=\frac{(1-o_t)^{\beta_j-1}}{b_{q_t}(o_t)}\left[\frac{\partial {o_t}^{\alpha_j-1}}{\partial \alpha_j}\cdot\frac{1}{\mathrm{B}(\alpha_j,\beta_j)}+{o_t}^{\alpha_j-1}\cdot\frac{\partial}{\partial \alpha_j}\left(\frac{1}{\mathrm{B}(\alpha_j,\beta_j)}\right)\right]$$
$$=\frac{(1-o_t)^{\beta_j-1}}{b_{q_t}(o_t)}\left[\frac{{o_t}^{\alpha_j-1}\log o_t}{\mathrm{B}(\alpha_j,\beta_j)}+{o_t}^{\alpha_j-1}(-1)\left(\frac{1}{\mathrm{B}(\alpha_j,\beta_j)}\right)^2(\psi(\alpha_j)-\psi(\alpha_j+\beta_j))\mathrm{B}(\alpha_j,\beta_j)\right]$$
$$=\frac{{o_t}^{\alpha_j-1}(1-o_t)^{\beta_j-1}}{\mathrm{B}(\alpha_j,\beta_j)b_{q_t}(o_t)}\left[\log o_t-\psi(\alpha_j)+\psi(\alpha_j+\beta_j)\right]=\frac{b_{q_t}(o_t)}{b_{q_t}(o_t)}\left[\log o_t-\psi(\alpha_j)+\psi(\alpha_j+\beta_j)\right]$$
$$=\log o_t-\psi(\alpha_j)+\psi(\alpha_j+\beta_j)$$

$$\frac{\partial\left(\log b_{q_t}(o_t)\right)}{\partial \beta_j}=\frac{1}{b_{q_t}(o_t)}\frac{\partial\left(b_{q_t}(o_t)\right)}{\partial \beta_j}=\frac{1}{b_{q_t}(o_t)}\cdot\frac{\partial}{\partial \beta_j}\left(\frac{{o_t}^{\alpha_j-1}(1-o_t)^{\beta_j-1}}{\mathrm{B}(\alpha_j,\beta_j)}\right)$$
$$=\frac{{o_t}^{\alpha_j-1}}{b_{q_t}(o_t)}\left[\frac{\partial(1-o_t)^{\beta_j-1}}{\partial \beta_j}\cdot\frac{1}{\mathrm{B}(\alpha_j,\beta_j)}+(1-o_t)^{\beta_j-1}\cdot\frac{\partial}{\partial \beta_j}\left(\frac{1}{\mathrm{B}(\alpha_j,\beta_j)}\right)\right]$$
$$=\frac{{o_t}^{\alpha_j-1}}{b_{q_t}(o_t)}\left[\begin{array}{l}\dfrac{(1-o_t)^{\beta_j-1}\log(1-o_t)}{\mathrm{B}(\alpha_j,\beta_j)}+(1-o_t)^{\beta_j-1}\\ \times(-1)\left(\dfrac{1}{\mathrm{B}(\alpha_j,\beta_j)}\right)^2(\psi(\beta_j)-\psi(\alpha_j+\beta_j))\mathrm{B}(\alpha_j,\beta_j)\end{array}\right] \qquad \text{(A.7)}$$
$$=\frac{{o_t}^{\alpha_j-1}(1-o_t)^{\beta_j-1}}{\mathrm{B}(\alpha_j,\beta_j)b_{q_t}(o_t)}\left[\log(1-o_t)-\psi(\beta_j)+\psi(\alpha_j+\beta_j)\right]$$
$$=\frac{b_{q_t}(o_t)}{b_{q_t}(o_t)}\left[\log(1-o_t)-\psi(\beta_j)+\psi(\alpha_j+\beta_j)\right]$$
$$=\log(1-o_t)-\psi(\beta_j)+\psi(\alpha_j+\beta_j)$$

Note that $\psi(\bullet)$ denotes digamma function formed during differentiation of beta function $\mathrm{B}(\alpha_j, \beta_j) = \Gamma(\alpha_j)\Gamma(\beta_j) / \Gamma(\alpha_j + \beta_j)$. The derivatives of beta function are determined as follows:

$$\frac{\partial \mathrm{B}(\alpha_j, \beta_j)}{\partial \alpha_j} = \mathrm{B}(\alpha_j, \beta_j)\left[\frac{\Gamma'(\alpha_j)}{\Gamma(\alpha_j)} - \frac{\Gamma'(\alpha_j + \beta_j}{\Gamma(\alpha_j + \beta_j)}\right] = \mathrm{B}(\alpha_j, \beta_j)\left[\psi(\alpha_j) - \psi(\alpha_j + \beta_j)\right] \quad \text{(A.8)}$$

$$\text{Also,} \quad \frac{\partial \mathrm{B}(\alpha_j, \beta_j)}{\partial \beta_j} = \mathrm{B}(\alpha_j, \beta_j)\left[\psi(\beta_j) - \psi(\alpha_j + \beta_j)\right]$$

Therefore, the emission density parameter estimation problem reduces to solution of following two equations:

$$\begin{aligned} &\sum_{j=1}^{K}\sum_{t=1}^{T} P(O, q \mid \mathrm{M}')\left[\log(o_t) - \psi(\alpha_j) + \psi(\alpha_j + \beta_j)\right] = 0 \\ &\sum_{j=1}^{K}\sum_{t=1}^{T} P(O, q \mid \mathrm{M}')\left[\log(1 - o_t) - \psi(\beta_j) + \psi(\alpha_j + \beta_j)\right] = 0 \end{aligned} \quad \text{(A.9)}$$

The parameter estimation procedure outlined above was repeated, the log-likelihood increased with every iteration, until the solutions for different unknowns converged. The forward-backward algorithm computations for large datasets involved summation of a large number of terms that exceeded the precision range of computing machines. However, these steps are inevitable for estimation of parameters in HMM. In order to cope with this issue, scaling was performed [Rabiner, 1989].

The posterior probability of being in a particular drought state at time t, that forms the basis for estimating the uncertainty in drought state classification, is given by:

$$P(q_t = i \mid O, M') = \frac{\alpha^*_t{}'(i)\beta^*_t{}'(i)}{\sum_{i=1}^{K} \alpha^*_t{}'(i)\beta^*_t{}'(i)} \quad \text{(A.10)}$$

Appendix B

Tabulated Results at 3- and 6-month Time Scale

Table B1 HMM beta emission distribution parameters for different dry/wet states in the 3-month time scale drought classification models used in the study

Model	State 1				State 2				State 3			
	α	β	Mean	Variance	α	β	Mean	Variance	α	β	Mean	Variance
1	1	30	0.032	0.001	2	9	0.182	0.012	5	6	0.455	0.021
2	2	24	0.077	0.003	3	6	0.333	0.022	6	3	0.667	0.022
3	2	20	0.091	0.004	12	31	0.279	0.005	28	27	0.509	0.004
4	1	20	0.048	0.002	3	8	0.273	0.017	10	5	0.667	0.014
5	1	30	0.032	0.001	3	16	0.158	0.007	4	7	0.364	0.019
6	1	22	0.043	0.002	3	14	0.176	0.008	7	9	0.438	0.014
7	1	24	0.04	0.001	3	11	0.214	0.011	10	9	0.526	0.012

Model	State 4				State 5			
	α	β	Mean	Variance	α	β	Mean	Variance
1	18	5	0.783	0.007	42	2	0.955	0.001
2	21	2	0.913	0.003	122	2	0.984	0.000
3	34	12	0.739	0.004	20	2	0.909	0.004
4	19	2	0.905	0.004	156	2	0.987	0.000
5	20	9	0.69	0.007	28	3	0.903	0.003
6	20	8	0.714	0.007	38	4	0.905	0.002
7	17	6	0.739	0.008	41	5	0.891	0.002

Table B2 HMM beta emission distribution parameters for different dry/wet states in the 6-month time scale drought classification models used in the study

Model	State 1				State 2				State 3			
	α	β	Mean	Variance	α	β	Mean	Variance	α	β	Mean	Variance
1	3	47	0.06	0.001	13	49	0.21	0.003	19	29	0.396	0.005
2	2	48	0.04	0.001	3	14	0.176	0.008	5	7	0.417	0.019
3	2	20	0.091	0.004	15	36	0.294	0.004	29	28	0.509	0.004
4	1	24	0.04	0.001	3	12	0.2	0.01	9	10	0.474	0.012
5	1	25	0.038	0.001	8	44	0.154	0.002	18	38	0.321	0.004
6	1	21	0.045	0.002	12	56	0.176	0.002	19	36	0.345	0.004
7	0.9	34	0.026	0.001	10	62	0.139	0.002	16	37	0.302	0.004

Model	State 4				State 5			
	α	β	Mean	Variance	α	β	Mean	Variance
1	21	12	0.636	0.007	14	2	0.875	0.006
2	15	6	0.714	0.009	23	2	0.92	0.003
3	34	14	0.708	0.004	18	2	0.9	0.004
4	19	6	0.76	0.007	28	2	0.933	0.002
5	26	20	0.565	0.005	13	3	0.813	0.009
6	28	20	0.583	0.005	17	4	0.81	0.007
7	30	27	0.526	0.004	16	5	0.762	0.008

Table B3 HMM transition probabilities for different dry/wet states in the 3-month time scale drought classification models used in the study

State	Model 1					Model 2					Model 3				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	0.59	0.41	0.00	0.00	0.00	0.60	0.40	0.00	0.00	0.00	0.84	0.16	0.00	0.00	0.00
2	0.06	0.63	0.31	0.00	0.00	0.11	0.61	0.28	0.00	0.00	0.11	0.69	0.20	0.00	0.00
3	0.00	0.18	0.65	0.17	0.00	0.00	0.31	0.54	0.15	0.00	0.00	0.18	0.64	0.18	0.00
4	0.00	0.00	0.29	0.58	0.13	0.00	0.00	0.50	0.40	0.10	0.00	0.00	0.23	0.60	0.17
5	0.00	0.00	0.00	0.41	0.59	0.00	0.00	0.00	0.79	0.21	0.00	0.00	0.00	0.24	0.76

State	Model 4					Model 5					Model 6				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	0.71	0.29	0.00	0.00	0.00	0.78	0.22	0.00	0.00	0.00	0.79	0.21	0.00	0.00	0.00
2	0.10	0.74	0.16	0.00	0.00	0.12	0.64	0.25	0.00	0.00	0.11	0.69	0.20	0.00	0.00
3	0.00	0.29	0.60	0.11	0.00	0.00	0.21	0.61	0.18	0.00	0.00	0.19	0.63	0.18	0.00
4	0.00	0.00	0.49	0.44	0.07	0.00	0.00	0.36	0.51	0.12	0.00	0.00	0.31	0.53	0.16
5	0.00	0.00	0.00	0.92	0.08	0.00	0.00	0.00	0.44	0.56	0.00	0.00	0.00	0.37	0.63

State	Model 7				
	1	2	3	4	5
1	0.82	0.18	0.00	0.00	0.00
2	0.09	0.77	0.14	0.00	0.00
3	0.00	0.28	0.48	0.23	0.00
4	0.00	0.00	0.69	0.04	0.27
5	0.00	0.00	0.00	0.69	0.31

Table B4 HMM transition probabilities for different dry/wet states in the 6-month time scale drought classification models used in the study

State	Model 1					Model 2					Model 3				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	0.70	0.30	0.00	0.00	0.00	0.63	0.37	0.00	0.00	0.00	0.89	0.11	0.00	0.00	0.00
2	0.13	0.65	0.21	0.00	0.00	0.10	0.60	0.30	0.00	0.00	0.08	0.82	0.10	0.00	0.00
3	0.00	0.21	0.56	0.23	0.00	0.00	0.16	0.70	0.14	0.00	0.00	0.13	0.69	0.18	0.00
4	0.00	0.00	0.19	0.67	0.13	0.00	0.00	0.19	0.70	0.11	0.00	0.00	0.17	0.72	0.11
5	0.00	0.00	0.00	0.16	0.84	0.00	0.00	0.00	0.23	0.77	0.00	0.00	0.00	0.14	0.86

State	Model 4					Model 5					Model 6				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	0.78	0.22	0.00	0.00	0.00	0.86	0.14	0.00	0.00	0.00	0.84	0.16	0.00	0.00	0.00
2	0.09	0.76	0.15	0.00	0.00	0.11	0.72	0.17	0.00	0.00	0.15	0.68	0.17	0.00	0.00
3	0.00	0.19	0.67	0.14	0.00	0.00	0.18	0.62	0.20	0.00	0.00	0.14	0.68	0.18	0.00
4	0.00	0.00	0.21	0.71	0.08	0.00	0.00	0.22	0.66	0.13	0.00	0.00	0.17	0.72	0.11
5	0.00	0.00	0.00	0.30	0.70	0.00	0.00	0.00	0.16	0.84	0.00	0.00	0.00	0.14	0.86

State	Model 7				
	1	2	3	4	5
1	0.84	0.16	0.00	0.00	0.00
2	0.15	0.63	0.21	0.00	0.00
3	0.00	0.19	0.61	0.20	0.00
4	0.00	0.00	0.23	0.64	0.13
5	0.00	0.00	0.00	0.17	0.83

VITA

VITA

MEENU RAMADAS
mramadas@purdue.edu

EDUCATION

Aug 2011-May 2015	**Doctor of Philosophy** Advisor : Dr. Rao S Govindaraju Lyles School of Civil Engineering, Purdue University, IN, USA Thesis Title: Probabilistic Models for Droughts: Applications in Trigger Identification, Predictor Selection and Index Development
Aug 2009-May 2011	**Master of Engineering** Water Resources and Environmental Engineering Advisor: Dr. Pradeep Mujumdar Indian Institute of Science Bangalore, Karnataka, India Thesis Title: Hydrological Modelling and Assessment of Impact of Climate Change: Case Study of the Tungabhadra Basin
Aug 2005-May 2009	**Bachelor of Technology** Civil Engineering Advisor: Dr. Ajitha Bhaskar College of Engineering, Trivandrum (Kerala University), India Thesis Title: Tensile and Pull Out Tests on Coir Geotextiles

HONORS AND AWARDS

2015	The Estus H. and Vashti L. Magoon Award Awarded to graduate students for excellence in teaching, by the College of Engineering, Purdue University.
2014, 2012	Jacques W. Delleur Award Awarded to graduate students doing research on hydraulics, hydromechanics, surface or ground water hydrology, or water resources engineering, in Purdue University.
2009	All India Rank 49 in Graduate Aptitude Test in Engineering Conducted by Ministry of Human Resources Development, Government of India, for post-graduate engineering admissions.

TEACHING EXPERIENCE

Spring 2014,2015	Course Instructor, Purdue University Course Title: CE 343 – Hydraulics Lab
Spring 2014	Teaching Assistant, Purdue University Course Title: CE 340 – Hydraulics taught by Dr. Dennis Lyn

RESEARCH EXPERIENCE

2011-Present	Research Assistant (Doctoral level), Purdue University
2009-2011	Research Assistant (Master's Level), IISc Bangalore
2008-2009	Research Assistant (Undergraduate level), Kerala University

PEER-REVIEWED PUBLICATIONS

1. Ramadas, M., Ojha, R., and Govindaraju, R. (2015). "Current and future challenges in groundwater II. Water quality modeling." Journal of Hydrologic Engineering, 20, SPECIAL ISSUE: Grand Challenges in Hydrology, A4014008. doi:10.1061/(ASCE)HE.1943-5584.0000936.

2. Ojha, R., Ramadas, M., and Govindaraju, R. (2015). "Current and future challenges in groundwater I. Modeling and management of resources." Journal of Hydrologic Engineering, 20, SPECIAL ISSUE: Grand Challenges in Hydrology, A4014007. doi:10.1061/(ASCE)HE.1943-5584.0000928.

3. Ramadas, M., and Govindaraju, R. S. (2014). "Probabilistic assessment of agricultural droughts using graphical models." Journal of Hydrology. doi:10.1016/j.jhydrol.2014.09.026.

4. Ramadas, M., Maity, R., Ojha, R., and Govindaraju, R. S. (2014). "Predictor selection for streamflows using a graphical modeling approach." Stochastic Environmental Research and Risk Assessment. doi: 10.1007/s00477-014-0977-1.

5. Maity, R., Ramadas, M., and R. S. Govindaraju (2013). "Identification of hydrologic drought triggers from hydroclimatic predictor variables." Water Resources Research, 49, 4476-4492. doi:10.1002/wrcr.20346.

6. Meenu, R., Rehana, S., and Mujumdar, P. P. (2013). "Assessment of hydrologic impacts of climate change in Tunga-Bhadra river basin, India with HEC-HMS and SDSM." Hydrological Processes, 27(11), 1572–1589. doi: 10.1002/hyp.9220.

CONFERENCES

1. Ramadas, M., and Govindaraju, R. S. (2014) A new index for agricultural droughts based on crop needs and available soil moisture. Abstract submitted to HYDRO 2014 International, Paper Id: HYDRO2014_194, MANIT, Bhopal, India. Dec. 18-20, 2014.

2. Ramadas, M., Chaubey, I., Niyogi, D., Song, C. X., and Govindaraju, R. S. (2014) Probabilistic assessment of agricultural droughts using graphical models. In 69th SWCS International Annual Conference, Lombard, IL. July 27-29, 2014.

3. Ramadas, M., and Govindaraju, R. S. (2014) Probabilistic assessment of agricultural droughts using graphical models. In 2014 Hydro-Climate Symposium on Modeling Climate Change, Abstract Id: 478, 2014 World Environmental and Water Resources Congress, Portland, OR. June 1-5, 2014.

4. Ramadas, M., Maity, R., Chaubey, I., Niyogi, D., Song, C. X., Nendunuri, K. V., and Govindaraju, R. S. (2013) Identification of hydrologic drought triggers from hydro-climatic predictor variables. Poster presented at the 2013 SWCS International Annual Conference, Reno, NV. July 21-24, 2013.

5. Ramadas, M. and Govindaraju, R.S. (2013) Unsaturated flow in vertically non-uniform soils: derivation of sharp front models for infiltration and redistribution. In 11th Symposium on Groundwater Hydrology, Quality and Management, Abstract Id: 382, 2013 World Environmental and Water Resources Congress, Cincinnati, OH. May 19-23, 2013.

6. Ramadas, M., Maity, R., and Govindaraju, R. S. (2012) Dimensionality reduction in hydro-climatic variables for probabilistic streamflow prediction using a hybrid approach. In Hydro-climate Symposium, Abstract Id: 632, World Environmental and Water Resources Congress, Albuquerque, NM. May 20-24, 2012.

7. Ramadas, M., and Mujumdar, P. P. (2012) Assessment of hydrologic impacts of climate change in Tunga-Bhadra River Basin, India with HEC-HMS and SDSM. Poster presented at the 2012 World Environmental and Water Resources Congress, Albuquerque, NM. May 20-24, 2012

8. Chaubey, I., Ramadas, M., Mallya, G., Ojha, R., Govindaraju, R. S., Niyogi, D., Song, C. X., and Nendunuri, K.V. (2012) Development of drought triggers for agricultural applications. Poster presented at the 2012 Land Grant and Sea Grant National Water Conference, Portland, OR. May 20-24, 2012.

PROFESSIONAL MEMBERSHIPS

2014-Present	Student Member, American Society of Civil Engineers (ASCE)
2013-Present	Student Member, Soil and Water Conservation Society (SWCS)
2012-Present	Member, Purdue Water Community

ACTIVITIES AND INTERESTS

Volunteer : ASHA for Education Purdue Chapter

Hobbies : Poetry writing, Reading, Traveling, Designing

Games : Badminton, Table Tennis