An assessment of the onset of summer rainy season in Southern Africa - case study of Botswana.

Denis C. Cheruiyot

A MINI DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT FOR THE AWARD OF A MASTER OF PHILOSOPHY IN INFORMATION TECHNOLOGY,

UNIVERSITY OF CAPE TOWN, SOUTH AFRICA

ABSTRACT

The economies of most Sub-Saharan African countries are linked to the onset, reliability and performance of seasonal rainfall. Failure of seasonal rains may signal food deficits or worse. Farmers, water conservationists and government bodies responsible for food security, all have an interest in seasonal rainfall: onset, approximate dates for start of the season and probabilities for early, normal or late onset of rains. This knowledge enables them make crucial decisions as to the choice of crops, planting dates, management of dams, pasture and hydro-electric dams.

In this thesis, daily rainfall data for 29 rainfall stations in Botswana for the years 1971 - 2004 was analyzed to determine Start-of-Season (SoS)/ Onset of summer rainfall. We used Principal Component Analysis to determine rainfall homogeneous zones in Botswana. Basically three regions were identified for October, November December (OND) rainfall months. Rainfall values in representative stations in each zone (Northern, Central and South-Eastern and Western regions) were correlated with Sea Surface Temperatures (SSTs) in global oceans to determine ocean regions that correlate well with Botswana rainfall.

The onset dates were grouped into *false*, *early*, *normal*, *late* and *failed* onsets. Monthly rainfall and Rainfall Onsets for selected 14 rainfall stations and ten other weather parameters, (that include SSTs, Sea Level Pressures (SLPs) and climate indices) were placed in a spreadsheet. Emergent Situation Awareness (ESA) for dynamic Bayesian networks (DBN) was used to analyze this data. The ESA for DBN models temporal dependencies among the weather parameters and climate indices using Direct Acyclic Graphs (DAG). This innovative DBN technology, ESA, reveals more detailed information from complex models. It reveals what is currently happening over time in a domain of interest. Each of the parameters and climate indices revealed varying degrees of beliefs for early, normal, late or failed rainfall onsets in Botswana. Some of the parameters which showed higher degrees of beliefs are promising signals to the onset of summer rains.

Acknowledgments

I am grateful to the Lord Almighty who enabled me undertake this work.

I am thankful to my supervisor, Dr Anet Potgieter for her guidance in this work. I am also thankful to Mr. Isaac Osunmakinde (PhD student, Computer Science, University of Cape Town) for his valuable support and assistance with programming experiments.

I am also grateful to my former colleagues at Botswana Meteorological Services, especially Mr. Isaac Kusane and Mr. Samuel Machua of Kenya Meteorological Department for their assistance with data collection and sourcing appropriate software for use in this work.

I will miss Mr. Blessing Siwela of SADC Early Warning Unit, Gaborone, Botswana, who was a fellow student in this course. In the absence of face to face interaction with lecturers, he was an encouragement and inspiration through uncertainty and challenging times in this course that was in most aspects an Online/Distance Course.

Many thanks and much credit goes to my lovely wife, Mary and daughters Karen, Sharon and Gloria for their support and encouragement to complete this work.

Table of Contents

List of Figures

List of Tables

Glossary of Terms

SLP	Sea Level Pressure
SST	Sea Surface Temperature
hPa	Hecto-Pascal
WMO	World Meteorological Organization
NCAR	National Center for Atmospheric Research
NCEP	National Center for Environmental Prediction
ECMWF	European Centre for Medium-range Weather Forecast
ERA-40	ECMWF 40 Year Re-Analysis
NOAA	National Oceanic and Atmospheric Administration
FEWS	Famine Early Warning System
FAO	Food and Agricultural Organization
SADC	South African Development Community
USAID	United States Agency for International Development
CPC	Climate Prediction Center
SOS	Start of Season

ITCZ	Inter Tropical Convergence Zone
SWIO	South West Indian Ocean
LTM	Long Term Mean
TTT	Temperate Tropical Trough
SOI	Southern Oscillation Index
BN	Bayesian Network
DBN	Dynamic Bayesian Network
CPT	Conditional Probability Table
Pdf	Probability Distribution Function
DAG	Directed Acyclic Graph
AgroMet	Agricultural-Meteorology
PCA	Principal Component Analysis
ESA	Emergent Situation Awareness
TAO	Tropical Atmosphere/Ocean
TRITON	Tropical Pacific Ocean Tele-metering Oceanographic and Meteorological data

Chapter 1: Introduction and Motivation

As in most parts of tropical Africa, the date of the onset of rains is of major importance mainly to farming communities and water conservationists in Southern Africa. In most African countries with small economies like Botswana, rainfed agriculture is important (Todd, Washington & Palmer, 2004), (Umoh, 2006).

While the summer rainy season for much of Southern Africa generally lasts from September to April, rains occur in Botswana mainly from November - March (Washington & Todd 1999). There is very little rainfall during winter months (May - September). The months of October, November, December (OND) can be considered as onset months while January, February, March (JFM) as cessation months. Summer rainfall is crucial to agricultural activities, management of pasture, water conservation and hydro-electric dams.

Planting of important crops such as maize, sorghum and wheat depend on temporal and spatial distribution of rainfall, and on the onset dates. Early planting may lead to crop failure, while late planting may also reduce crop yields due to rainfall deficits, water logging or reduced rainfall duration.

Botswana is a landlocked country, (see figure 1). It occupies an area of approximately 580000 km^2 and lies between latitudes 17.5°S to 27.5°S and longitudes 20°E and 29°E. Much of the country is a plateau with elevation ranging from 700m - 1300 m. The highest point in the country (1491m) is situated in hilly area in South-east near Otse-Kanye and Molepolole areas.

Botswana, like her neighboring countries, has an arid to semi-arid climate with large annual rainfall variability (Mason & Jury 1997). From available data, the mean annual rainfall varies from a minimum of less than 250mm in Kgalagadi District (extreme South-West) to a maximum of 650mm in the extreme North-East (around Kasane). There is a secondary maximum of over 550mm around Lobatse in the southern border with South Africa and secondary minimum around Baines Drift along the lower Limpopo River, to the extreme north-eastern tip of the country, See Figures 2a and 2b.

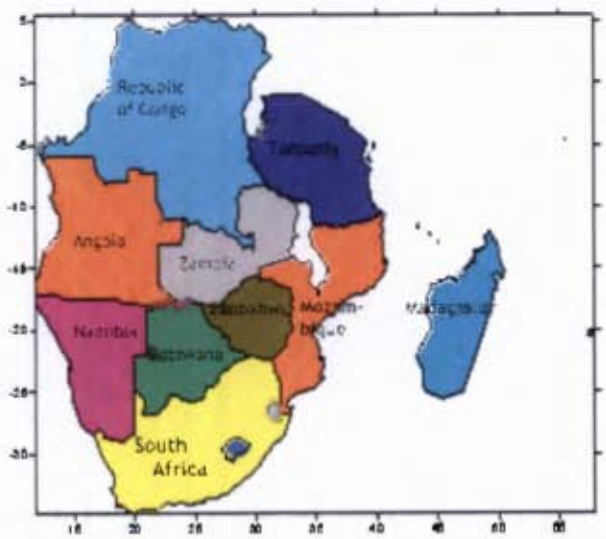


Figure 1: Southern African countries

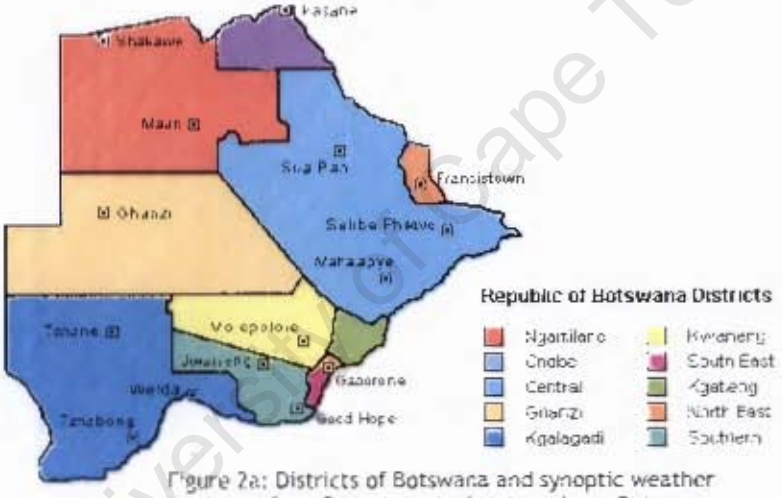


Figure 2a: Districts of Botswana and synoptic weather stations - from Department of Meteorology, Botswana website - http://www.weather.info.bw/

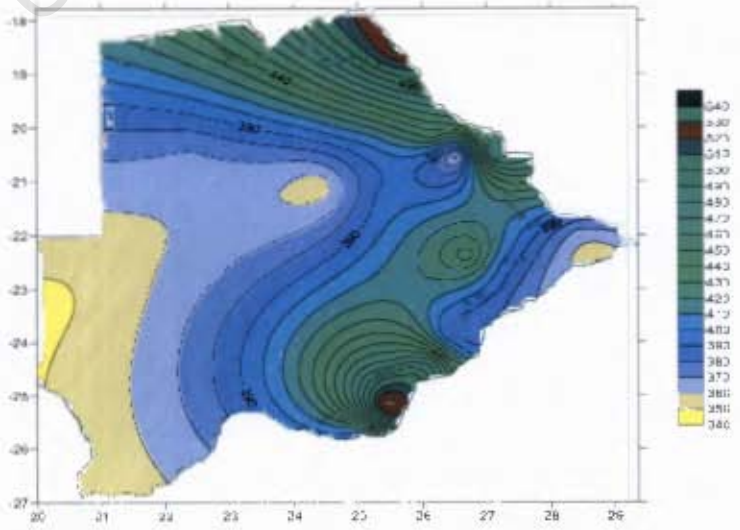


Fig.2b. Mean annual (mm) rainfall in Botswana

1.1 Motivation

Onset of seasonal rains, referred also as Start-of-Rains or Start-of-Season (SoS), does not have a unique definition (Ati, Stigter & Oladipo, 2002). In much of West Africa, the onset of seasonal rains is linked to the latitudinal shift of the Inter-Tropical Convergence Zone (ITCZ) from 5°N in May/June to another quasi-stationary location at 10°N in July/August, (Ati et al., 2002). The convergence line between dry North-easterly Harmattan winds and moist south-westerly monsoons bring in sufficient moisture that trigger convective systems along the ITCZ.

The Famine Early Warning System (FEWS) program of the United States Agency for International Development monitors and forecasts the productivity of crops that are unique to Sahel (West Africa), the Greater Horn of Africa (East Africa) and Southern African regions. This information is available on their website http://edcintl.cr.usgs.gov/fews/wrsi.html. FEWS have developed a Water Requirements Satisfaction Index (WRSI) model, which is a function of rainfall estimate, potential evapo-transpiration and Start-of-Season. Start-of-Season is one of their products. The Start-of-Season is determined using a threshold amount and distribution of rainfall received in three consecutive dekads. In practice, FEWS reckons that the Start-of-Season "*is generally established when there is at least 25mm of rainfall in one dekad, followed by a total of at least 20 mm of rain in the next two dekads*". Failed planting is declared if in the following two dekads, rainfall received does not total 20 mm.

Definition and classification of onset of rains used in this work is given in section 4.3. A *mean onset date* was determined for the whole country. Subjective criteria were used to classify onset dates into *false, early, normal, late and failed* onsets. A *normal* onset of rainfall is one that occurred on dates within a 20 day period around the mean onset date, that is, 10 days before and 10 days after the *mean* onset date. An *early* onset is one that occurred within a 20 day period, earlier than earliest normal onset. A *late* onset is one that occurred within 20 day period later than the latest normal onset date. Any onset event later than *late* onset was considered a *failed* onset, while an onset event earlier than *early* onset was considered *false* onset.

and climate indices. The technology will output Directed Acyclic Graphs (DAGs) and Probabilities that will represent dependencies among weather parameters/climatic indices and summer rainfall onsets in Botswana.

This thesis will address the following questions:

- Are farmers, water conservationists and government departments responsible for food security able to know the probabilities for false, early, normal, late or failed onset of rainfall, given various atmospheric conditions and weather parameters?
- What are the basic indicators for the onset of rains in Southern Africa sub-continent?

This study aims to determine degrees of beliefs and probabilities for false, early, normal, late or failed onset of rains given *normal* and *anomalies* of weather parameters and climate indices. For example, what is the probability of a normal onset of rains given an anomalously warm sea surface temperature in the Indian Ocean? Is it highly probable that we shall have early onset of rains given a cooler Atlantic Ocean and positive anomalies of sea level pressure?

This main objective of this study is to examine the weather parameters and climatic indices that are the best indicators for the onset of summer rains in Southern Africa using the case study of Botswana.

Chapter 2: Atmospheric features which influence weather in Southern Africa

2.0 Introduction

There are a number of distinct large scale atmospheric weather patterns and synoptic systems that influence rainfall and onset of rainfall in Southern Africa. Some of these are El-Niño/Southern Oscillation (ENSO), Sea Surface Temperatures (SSTs), Sub-tropical High Pressure systems, synoptic scale wind patterns and position of the Inter-Tropical Convergence Zone (ITCZ), (Washington & Todd, 1999). Others are extra-tropical cyclones and easterly/westerly waves.

A broad description of these features is included in this chapter. Each of these systems has an influence on rainfall and the onset dates of rainfall in the region in varying degrees.

2.1 El-Niño/La-Niña

El-Niño, a Spanish term for "the boy", is a term that is associated with anomalous warming of the sea-surface off the coast of Peru, in East Pacific. This is an oceanic area that normally has relatively cool water all year round. The event is variously referred as a "warm event" or a "warm episode". The opposite episode, *La-Niña* - "the girl" is the cooling of sea-surface temperatures in the equatorial Pacific Ocean which influences atmospheric circulation, and consequently rainfall and temperature in specific areas around the world. It is referred to as a "cold event".

The Southern Oscillation is an east-west balancing movement of air masses between the Pacific and Indo-Australia regions. It is associated with trade wind systems and El-Niño. The process is quantified by the Southern Oscillation Index (SOI). The two processes, the oceanic component, El-Niño, and the other in the atmosphere, the Southern Oscillation, interact to form what is referred to as the El-Niño/Southern Oscillation or ENSO. The anomalous warming of sea surface waters in East Pacific and consequent shifting of convective zones from Indonesia to Central Pacific causes a reversal of east-west tropospheric circulations that affect weather patterns globally (Wang, 2002).

The variation of pressure between Tahiti (Central Pacific) and Darwin (Northern Australia) has been used to create the Southern Oscillation Index (SOI). Generally large negative pressure values are associated with warm events (*El-Niño*), while large positive pressure values are associated with cold events (La-Niña). (Current information and updates of ENSO can be obtained from National Oceanic and Atmospheric Administration's Climate Prediction Center (CPC) website: *http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensocycle/soi.shtml*). Figure 3 shows a time series of *standardized* SOI.

Rainfall variability in southern Africa exhibits a strong relationship to the ENSO cycle (Nicholson et al., 2001). Early in the season, October through to December, a warm El-Niño episode is associated with below normal rainfall conditions in Southern Africa. Further north, in Eastern Africa, the El-Niño phenomenon is associated with above normal rains. The situation is reversed during the La-Niña episode where Southern Africa experiences anomalously wet conditions and above normal rainfall (Nicholson & Selato 2000), (Semazzi & Indeje, 1999).

In Southern Africa, and specifically Botswana, below normal rains and severe drought conditions have been associated with ENSO, (Nicholson et al. 2001), (Jury, 2002). As already mentioned, the strong El-Niño events of 1982/03 La-Niña and 1991/92 led to severe droughts in Southern Africa (Jury, 2002), and Botswana in particular (Food and Agricultural Organization (FAO), 2004)). From available data, in 1982 most rainfall stations in Botswana received about 85% of the annual long term mean (LTM) rainfall; in 1983/84 most stations recorded 70% of LTM; in 1992, stations received 58% of LTM and in 1993, stations received 69% of LTM. In 1992 Sehitwa and Palapye stations both recorded annual historic lows of only 110 mm of rain.

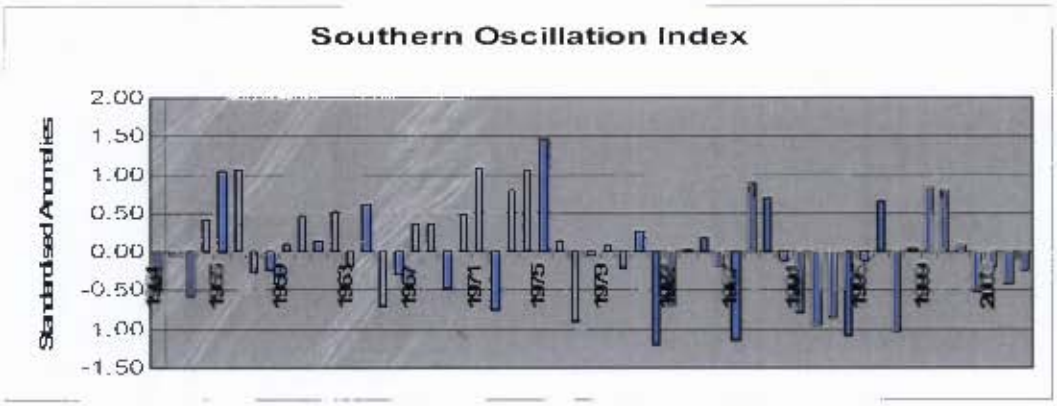


Figure 3: Time series of Southern Oscillation Indices (SOI).
(SOI = Tahiti (Central Pacific) SLP - Darwin (Northern Australia) SLP)

However not all El-Niño events in the past have led to massive rainfall deficits (Nicholson et al., 2001). The warm event of 1997 that extended to early 1998 has been billed as the El-Niño of the century. The episode conditions were accompanied by strong negative phase of Southern Oscillation, comparable in magnitude to that observed during 1982/83, see Figure 3. The event did not affect the Botswana rainfall significantly. The country did receive normal rains in both 1997 and 1998, on average recording 116.8% and 102% of the long term means respectively. Earlier on the 1986/87 El-Niño event also did not have much impact on rainfall in Botswana as the country received near normal rains. The Central and South Eastern regions of Botswana experienced a good 1996/97 rainy season with totals averaging 500 - 600mm (over 100mm above average). Nicholson et al. (2001) pointed out that in ENSO events that did not produce drought in Botswana, the SOI tended to be weak, with less consistent warming in Central and Western Pacific. This pattern leads to weaker atmospheric response with little or no influence on SSTs in Tropical Indian and Atlantic Oceans, which have an influence in Southern Africa sub-continent including Botswana (Nicholson et al.,2001), (Reason & Mulenga, 1999).

An example of a strong "warm event" is shown in Figure 4a. It is a NCEP/NCAR Reanalysis composite plot of SST anomalies for the months September - November, 1982. The base period is 1968-1996. We can see clearly the anomalous warming of positive 2°C to 3°C in the East Pacific, spreading to Central pacific. This was a pointer to the strong El-Niño episode of period 1982/83.

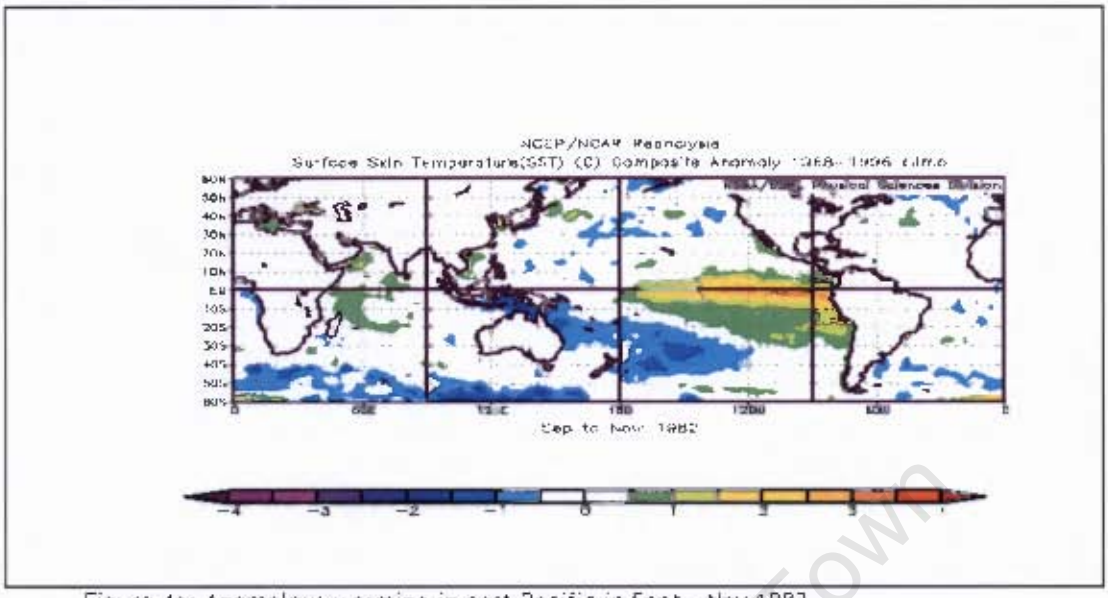


Figure 4a: Anomalous warming in east Pacific in Sept - Nov 1982

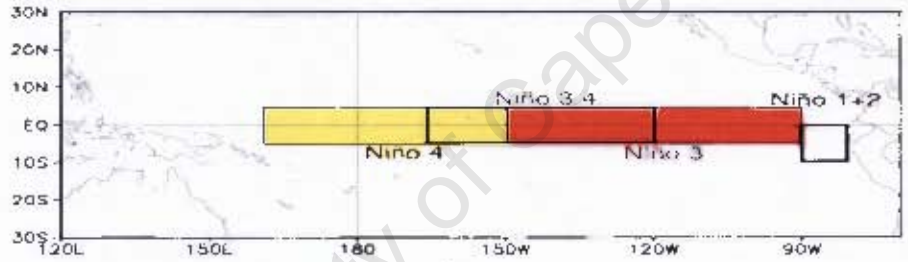


Figure 4b: Graphical depiction of four "NIÑO regions" along Central Pacific: from NOOA Climate PredictionCenter. Available from http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/Niño_regions.shtml

2.2 Sea Surface Temperatures

The sources of moisture and patterns of moisture transport to Southern Africa are the Sub-tropical High pressure system (Mascarene High) over South West Indian Ocean (SWIO) and St Helena High situated over South Atlantic Ocean (Washington & Preston, 2006). These features influence wind flows and moisture influx into the interior of Southern Africa. See Figure 5 below.

It is believed that anomalous SST gradients during summer in surrounding Indian and Atlantic oceans are a primary influence on rainfall patterns over Southern Africa (Nicholson & Selato, 2000), (Washington & Preston, 2006). These SSTs in the surrounding oceans exert influence on Southern African rainfall independent of ENSO (Reason, 2002). According to Klein (1999), whenever there is anomalous

heating or cooling of the ocean, for example during an ENSO year, it would take approximately 3-6 months, for corresponding changes in atmospheric circulations. Consequently Tropical North Atlantic and Indian Oceans would respond at least 3 months after onset of the event.

During ENSO years when El-Niño's influence on late summer rainfall (February to April) in Botswana is prominent, moisture transport is reduced over Southern Africa and shifts northwards into northern Mozambique and Malawi, leaving the southern regions drier (Nicholson et al. 2001). ENSOs influence on rainfall in Southern Africa is related to its influence on SSTs in Indian and Atlantic Oceans near the African continent (Nicholson et al. 2001).

While (Reason & Mulenga (1999) pointed out that warmer SSTs in SWIO are associated with wetter conditions over central regions of Southern Africa and vice-versa, (Todd et al. 2004) have shown that the establishment of SST gradients in the SWIO induces circulations favorable for dry and wet conditions over Southern Africa. Whenever there is anomalous warming in the south and cooling to the north of SWIO, the result is an anti-cyclonic circulation favorable for wet conditions over Southern Africa and vice-versa. In the first instance there will occur a longer and stronger track of deep and moist easterly/south easterly winds blowing into the interior of Southern Africa.

2.3 ITCZ and Subtropical Highs

The Inter-tropical Convergence Zone (ITCZ) is a zone of convergence between north-easterly and south-easterly trade winds near the equator and generally fluctuates following the position of the sun. It is marked by deep convective bands of clouds and associated rains and thunderstorms. While ITCZ is well marked over Indian, Atlantic Oceans and Western Africa, its structure and position is rather complicated in East, Central and Southern Africa. Its characteristics are modified by varied topography and thermally induced meso-scale circulations near or around lakes, seas, rivers (Okoola, 1999).

The "meridional arm" of the ITCZ, sometimes referred by weathermen in the region as Congo Air Boundary (CAB), is a zone of convergence of south westerlies from the

Atlantic Ocean and easterlies from the Indian Ocean, (see Figure 5). In Southern Africa, the locations and strengths of the meridional and zonal arms of the ITCZ are controlled by the anticyclonic high pressure system in the South Atlantic (referred to as St Helena High), and another anticyclone south of Madagascar island (Mascarene High).

Summer rainfall in Southern Africa is received most consistently over the zonal stretch of ITCZ: Southern Congo, North-East Angola and Northern Madagascar, Usman & Reason (2004). Any further southward movement or fluctuation of the ITCZ could be associated jointly or with one of the following synoptic features: a cold front sweeping across South Africa; development of a low pressure area over Botswana or Zimbabwe; Temperate Tropical Trough (TTT) extending from South-East of South Africa to Angola/Namibia; and the presence of a tropical cyclone near Madagascar or parts of northern Mozambique channel. Moisture influx from the Congo and Southern Mozambique Channel supports the extended TTT (Makarau & Jury, 1997).

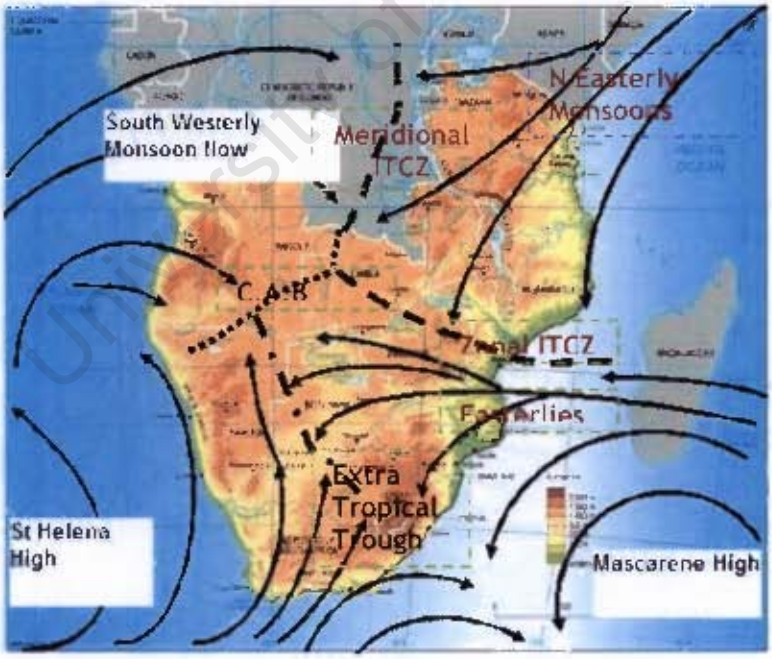


Figure 5: Synoptic features and air-masses over Southern Africa, after Tyson and Preston-Whyte, 2000. Background map taken from UNEP/GRID-Arendal Library of graphics resources.
http://maps.grida.no/go/graphic/southern_africa_topographic_and_political_map

2.4 Botswana Upper High

Botswana Upper High is a semi-permanent medium-level anticyclone found in Southern Africa, especially Namibia, Botswana and central and western South Africa that occur mainly during winter months (Preston-Whyte, 1999). This high pressure cell is found between levels (850 hPa[1] - 600 hPa), and occasionally occurs during summer months. This feature is characterized by large scale subsidence of air, significant atmospheric stability leading to suppression of convection, clear skies and most times prolonged dry spells. In summer months, these anticyclones sometimes cause prolonged heat waves. Preston-Whyte (1999) suggested that almost all droughts on the scale of dates, seasons or years may be linked to the predominance of this condition in Southern Africa.

2.5 Tropical/Extra-Tropical "Westerly" moving Troughs

Often called Temperate Tropical Trough (TTT), these troughs are inverted V-shaped North-West/South East orientating troughs in the westerly airstreams that occur in early and late summer, mainly found in Southern Africa and SWIO (Todd et al. 2004), (Washington & Todd, 1999). The passage of a deep trough and extensive *west to east* moving mid-latitude cyclone and associated cold front over the Cape and SE South Africa causes fall in pressure inland, and if there is favorable instability conditions and sufficient moisture, showers and thunderstorms may occur in the southern parts of Botswana.

These wave disturbances in the westerlies form an important link between extra-tropical features and the tropics and are associated with enhanced convection in the tropics (Todd et al.,2004), (Knippertz, 2007). On satellite images they appear as swats of white alto-cumulus or alto-stratus clouds. A deep westerly trough approaching the western coast of Namibia/Angola leads to falling pressures over the continent and deepening of surface ITCZ and consequently convective rainfall over parts of Botswana.

[1] hPa - hecto Pascal: A unit of measure of atmospheric pressure, used in Meteorology and Oceanography practice. Formerly units of measure used were the millibar (Mb). 1Mb=1hPa. 700hPa level is approximately equal to 3000 meters above sea level.

An example of an extra-tropical westerly wave and medium level conditions are depicted below. They are based on records from the Botswana Meteorological Service records for 13 August, 1980. There was a deep extra-tropical depression to the south east of Durban and associated cold and warm frontal systems. A deep trough extended inland to Botswana and South Western Zimbabwe.

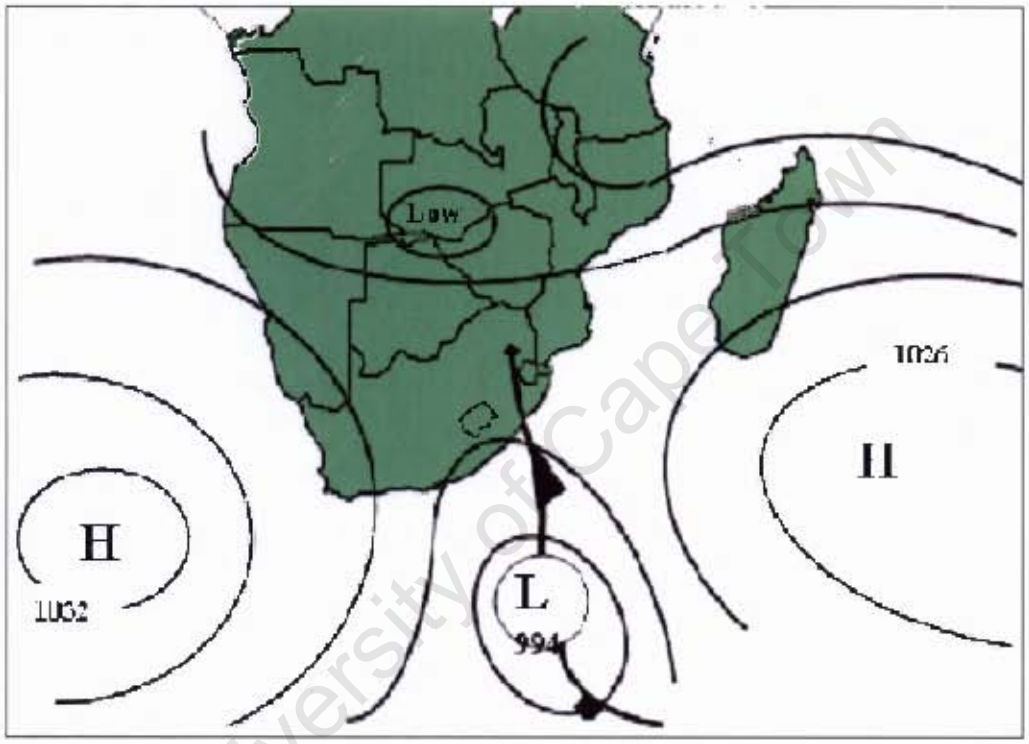


Figure 6a: A schematic representation of mean sea level surface synoptic conditions based on Botswana Department of Meteorology records for 13 August, 1980 1200 GMT.

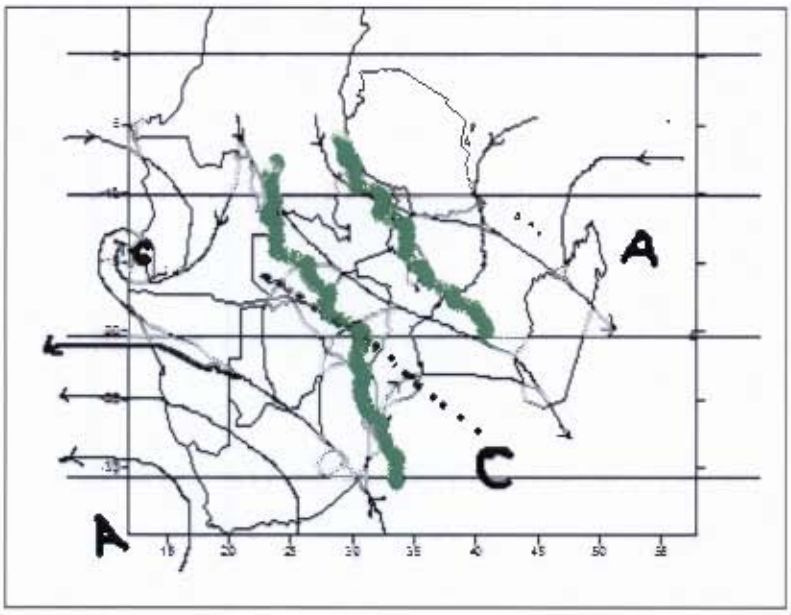


Figure 6b: Streamlines at 600 hPa on 18/08/1980. Rainfall area is bounded by green lines.
A - Anticyclonic flow
C- Cyclonic flow (around low pressure systems).
Trough line marked by dotted line

2.6 Tropical Cyclones

Tropical disturbances in the Indian Ocean form to the west of Australia, and are carried westward along the easterlies and recurve polewards around the Mascarene sub-tropical High in the South-West Indian Ocean. In late summer, between December and March, some of these tropical disturbances deepen to become tropical depressions; a few become full-fledged tropical cyclones (Preston-Whyte, 1999).

Fewer than 5% of fully developed tropical cyclones make on-shore landing in Mozambique and follow up the Limpopo River, (Reason & Keibel, 2004). The much publicized tropical cyclone "ELAINE" made a landing in Mozambique on 22nd February 2000, had winds in excess of 260 km/hr, traveled inland across Mozambique, Botswana and dissipated in South-East Namibia, (World Meteorological Organization (WMO), 2002). It caused massive flooding in Zambezi valley, eastern Zimbabwe and North-Eastern and South-Eastern Botswana. ELAINE killed hundreds of people and left tens of thousands homeless and caused damage that ran into millions of dollars (Reason & Keibel, 2004), (Du Plessis, 2002), (Jury, 2002), (World Meteorological Organization (WMO), 2002).

The impact of ELAINE was that between January and March (JFM) 2000, Baines-Drift in the North-Eastern tip of Botswana that normally receives 179mm JFM mean rainfall, recorded 696mm of rainfall; Martin's Drift recorded 510mm against LTM of 192mm and Kgale recorded 685mm against LTM of 260mm.

2.7 Conclusion

We have seen that large scale atmospheric features such as El-Niño/Southern Oscillation (ENSO), and synoptic scale systems such as the ITCZ, Tropical Cyclones, Sub-Tropical High Pressure systems influence the weather over Southern Africa. Their positions, timing and intensities determine the onset and performance of summer rains in Southern Africa.

Chapter 3: Background Theory on Bayesian networks

3.0 Introduction

We shall be considering several climate indices and weather parameters that may influence the onset of rains in Southern Africa such as SOI, SSTs, SLPs, 500 hPa geopotential height anomalies and 700 hPa winds. To model temporal decencies among these weather parameters using dynamic Bayesian networks, it is necessary to give background theory on Bayesian networks in this chapter.

We shall explore a dynamic Bayesian network (DBN) technology that was used in this work, called the Emergent Situation Awareness (ESA). We shall see how this technology evolves temporal models and reveals what is currently happening or change with time in a domain of interest.

3.1 Bayes' rule

Bayes' theorem relates to conditional and marginal probabilities of events, say A and B, and can be mathematically expressed as follows (Mihajlovic & Petkovic, 2001)

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)} \quad . \qquad (3.1)$$

Where: P(A) is the *prior* probability of A, it does not take into account any information about B;

P(A|B) is the conditional probability of A, given B, also called the *posterior* probability derived or depends on B;

P(B|A) is the *conditional* probability of B given A;

P(B) is the *marginal* probability of B, and acts as a normalizing constant.

Bayes' rule simply states:

$$Posterior = \frac{Conditional\ likelihood\ x\ Prior.}{Marginal\ Likelihood}$$

The rule can be modelled with the following equation, using *e* as *evidence* variable and H as *hypothesis*, (Mihajlovic & Petkovic, 2001):

$$P(H/e) = \frac{P(e/H)P(H)}{P(e)} \quad . \qquad (3.2)$$

The probability of a hypothesis, H, can be updated when evidence, e, has been obtained.

3.1.1 Bayesian Networks Definition

A Bayesian network is a graphical model for representing conditional independencies between a set of random variables in a probabilistic model, (Ghahrami, 1998). It is a specific type of graphical model which is a *directed acyclic graph (DAG)*, that is, all the edges of the graph are pointing at one particular direction *(directed)*. In this model there are also no *cycles*, one does not leave a vertex, move along an edge or a set of edges and arrive at the starting node.

A Bayesian network graph structure conforms to the following rules (Stephenson, 2000), (Murphy, 1998):

- The nodes or vertices represent the random variables, discrete or Continuous;
- Directed edges (arrows) between pairs of nodes, represent *cause* and *effect* relationships;
- The graph must be a Directed Acyclic Graph (DAG), it has no directed Cycles;
- The graph represents independence relationships between variables;
- Each node N has a set of conditional probability distributions P(N| Cause (N)) which can be stored in a conditional probability table (CPT). The table quantifies the effect of the cause(s) or parents on the Node N.

The Bayesian network illustrated in the figure below (Stephenson, 2000) consists of a set of nodes N = A, B, C, and a set of vertices or edges E = (B, A), (B, C). The edges (B, A) and (B, C) are directed from *parent* B to *children* A and C respectively. If C had a child of its own, say D, then B would be an *ancestor* of D

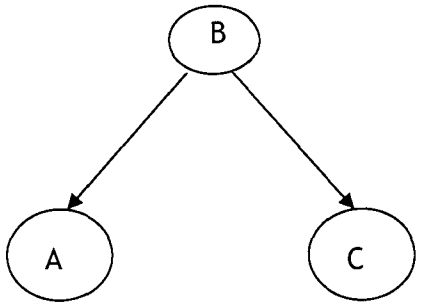


Figure 7: A simple Bayesian Network (probabilities are omitted), from Stephenson (2000)

Node A is dependent only on B, and hence P (A|B, C) = P (A|B).
Likewise, C is dependent only on B, and P (C|A, B) = P (C|B).

B is also independent from both A and C. Hence the joint probability distribution can be expressed as follows:

$$P(A, B, C) = P(A|B)\, P(B)\, P(C|B) \qquad (3.3)$$

The information in the Conditional Probability Table (CPT) for every node (attribute) defines a probability distribution that is used to predict node probabilities. Generally the probability of each node is

$$\text{node}|\text{ancestors}) = P(\text{node}|\text{parents}), \quad \text{(Stephenson, 2000).}$$

To generalize, for any set of random variables $X_i \mid i = 1{:}\ n$, the joint probability distribution over all variables is:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i|Pa(X_i)), \qquad (3.4)$$

where $Pa(X_i)$ are the parents of node X_i, (Heckerman, 1999), (Deviren & Khalid, 2001) .

Figure 8 below is an example of a Bayesian network, adopted from Neapolitan (2004). The Network represents probabilistic relationships among variables (entities of medical diagnosis of an individual with or without lung cancer). A history of smoking or not smoking may have a direct influence both on whether or not that individual has bronchitis, and on whether or not he/she has lung cancer. The

presence or absence of lung cancer has a direct influence on whether or not a chest x-ray is positive or negative.

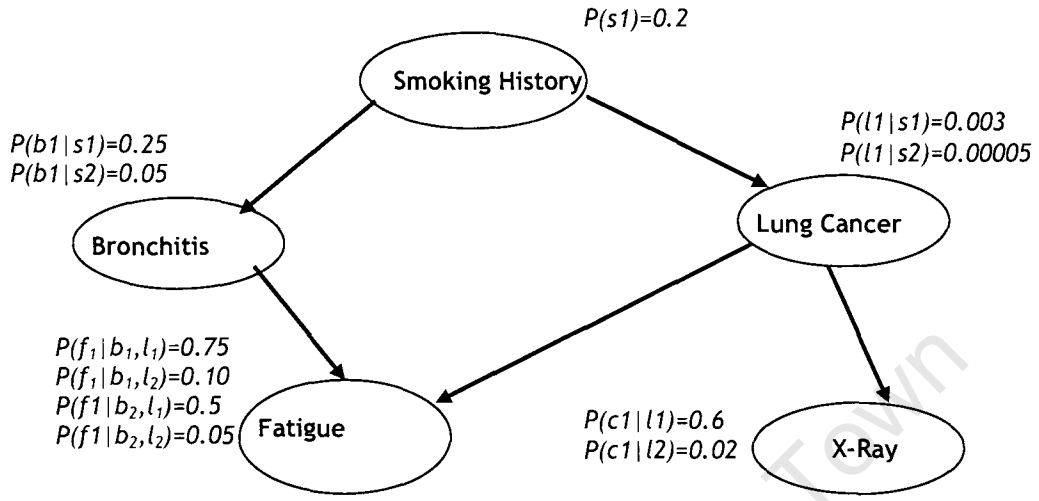


Figure 8: A Bayesian network for medical diagnosis of an individual, (Neapolitan, 2004)

The table next to each node represents the Condition Probability Table (CPT) for that particular node, and the table below is list of the nomenclature used in CPTs.

Table 1: List of nomenclature used in the conditional probability tables (CPTs)

s1 - There is a history of smoking	l1 - Lung cancer is present
s2 - There is no history of smoking	l2 - Lung cancer is absent
b1 - Bronchitis is present	f1 - Fatigue is present
b2 - Bronchitis is absent	f2 - Fatigue is absent
c1 - Chest X-ray is positive	c2 - Chest X-ray is negative

3.2 Bayesian Network Learning/Training

Two important features in a BN are the graph topology (structure) and parameters (data) of each conditional probability distribution. Table 2 below is a summary of four cases of learning structure and parameters, and methods for learning each case, suggested by (Murphy and Mian, 1999).

Table 2: Cases of learning structure and methods that are used

	Network Structure is..	Observability	Method
a	Known	Full	Use of Maximum Likelihood (ML) estimation algorithm.
b	Known	Partial	Use of Expectation Maximization (EM) algorithm or gradient ascent
c	Unknown	Full	Search through all possible network structures by using greedy algorithms (e.g. scoring metric) and clustering techniques to find possible clusters
d	Unknown	Partial, hidden or missing	Combination of the EM algorithm and a network structure search through model space.

In meteorology practice and atmospheric science, the state of the atmosphere can be quantified using various measurements. Parameters that can be measured or quantified include pressure, temperature, wind speeds and direction, rainfall, evaporation, sunshine and aspects of solar radiation. For a given situation, a BN structure may not be known, but the data variables are mostly observable, though in some cases data may not be available. For example, wind or rainfall may not have been observed in a given day; either the instrument was not working or the transmission was faulty. Hence we have cases where the structure of a BN is unknown and parameters are either fully observable and in some cases partially observable or missing.

3.2.1 Unknown structure, partial observability

In practice, in weather information, such as what has been described in this project, some of the data may be missing. So we not only lack complete data, but we also do not have a Bayesian network structure that can be easily distinguished. It means we have hidden nodes. One method to simplify this problem is the EM algorithm.

3.2.1.1 Expectation Maximization (EM) Algorithm Method

The EM algorithm was developed to compute the parameters of Bayesian networks and dynamic Bayesian networks (DBNs), and needs to be adjusted for learning BN

and DBN structure from incomplete data (Milhajlovic & Petkovic, 2001). Friedman (1998) proposed the *Structural Expectation maximization* (SEM) algorithm for use with EM algorithms to achieve the purpose.

The SEM algorithm has 'E' and 'M' steps. The 'E' step entails completing the data by computing expected counts based on the current structure and parameters. The 'M' step of SEM involves first recalculating *maximum likelihood* (ML), and secondly using the expected counts according to the second structure to evaluate other candidate structures. Maximum Likelihood is explained further in section 3.6.

Given the training data set D, and instances of all variables in the network $X(\{X_1, X_2, ..., X_N\})$ and let G represent a DAG whose vertices correspond to random variables of X. We are interested to find a Bayesian Network B = (G, θ) that best matches D. To evaluate a potential structure, one method is to compute the joint probability *P(D, G)*, for the data D and structure G, (Mihajlovic & Petkovic, 2001).

From Bayes' rule, we have $p(G,D) = p(D \mid G)p(G).$

Factorizing and taking its log, we have, from Mihajlovic & Petkovic, (2001)

$$\log\ p(D,G) = \log\ P(D \mid G) + \log\ p(G). \qquad (3.5)$$

One method to compute the posterior probability, $\log p(D,G)$, is to use an asymptotic approximation called the Bayesian Information Criterion (BIC) (Murphy & Mian, 1999)

$$\log\ P(D \mid G) \approx \log\ P(D \mid \theta'_G, G) - \frac{d}{2}\log\ M\ . \qquad (3.6)$$

Where

M is the number of samples (from the size of D);

θ'_G is the maximum likelihood estimate;

d is the dimension of the model, it is the number of free parameters in fully observable case.

The first term of the BIC score is the maximum likelihood of the parameters (computed using EM). It measures the model fit. The second term penalizes model complexity.

The BIC is a decomposable score; it decomposes into a sum of local terms, one per node. We still have computational complexity as we have to run EM at each step to compute θ'_G.

Friedman (1998) describes the SEM algorithm. One can describe learning both the structure and the variables as a combination of the methods and of adding new nodes, iterating as follows:

a. Add a new node to the network representing a hidden variable;
b. For this given set of nodes, find the best network connection;
c. Continue as long as the network keeps improving.

3.2.2 Unknown structure, full observability

When a number and type of some states in the network are known, but not their relation and mutual independence, then we would have to rely on expert knowledge and search for other ways to learn the structure of BN from observable data.

According to Heckerman (1999), one of the tasks that are needed to learn structure with complete training data is a metric for comparing potential structures against each other and a search algorithm for finding potential structures.

3.2.2.1 Score and Search Method

A local learning algorithm to learn a BN from a database of observations (rainfall and maximum wind speed) in a spatial network of 100 observing stations in Iberian Peninsula, Spain has been used by Cano, Sordo & Gutierrez (2004). For this case the network structure was not known but data variables were observable or derived. A "*score and search method*" was used to learn graphical structure and estimate probabilities from data.

The method computes the quality of the graphical structure and estimated parameters for the candidate BNs. The "search algorithm" component efficiently searches the space of possible BNs to find one with the highest quality.

To compute the quality of the graphical structure, for each BN, let B = (M, θ), where M is the Network structure and θ is the corresponding estimated probabilities. Let D be the available data. For each BN, assign a function of its posterior probability distribution P (B|D), calculated as follows, (Cano et al. 2004):

$$P(B \mid D) = P(M, \theta \mid D) = \frac{P(M, \theta, D)}{P(D)} \quad . \qquad (3.7)$$

$$\alpha\, P(M)P(\theta \mid M)P(D \mid M, \theta)$$

3.2.2.2 K2 Search Method

The K2 algorithm is a greedy search algorithm for finding a high quality BN in a reasonable time, (Cano et al. 2004). It uses the K2 metric for scoring networks. It assumes that all structures are considered equally likely. The algorithm assumes a complete database D of a sample of cases over a set of attributes that models the network.

Below is a modified K2 algorithm for graphical models used for learning a probabilistic model of rainfall, (Lee, 2006).

```
Input: Quantized data of n nodes, an ordering of n nodes, an ordering of neighbors for
       each node, max_parents
Output: Adjacency matrix representing all directed edges in the network
For i = 1 to n
parent_i = [ ]; #Initial condition: no parent node for any node
 P_old = f(i, parent_i); #Probability of data (i node) given parent_i
 Gonext = true;
  While Gonext & size(parent_i) < max_num
   P_new = f(i, parent_i, another parent_i); # choose from neighbors
    If P_new > P_old
     P_old = P_new;
     parent_i = parent_i + another parent_i;
    else Gonext = false;
  end
 Save parent nodes for node i; # in adjacent matrix
end
```

return adjacency matrix

The K2 algorithm is iterative and maximizes the K2 metric; it searches a set of parent nodes $\prod_i$ in every node. For each variable Y_i, the algorithm adds to its parent set $\prod_i$, the node that is lower numbered than Y_i and leads to maximum increment in the quality measure. Every node is enforced to have a maximum number of parents during the search.

A local K2 learning algorithm (or simply LK2) method that makes it more efficient in dealing with large number of variables (observing stations) and huge amounts of data has been suggested by Cano et al. (2004). The strategy was to modify sets of candidate parents of node Y_i {1,...., i-1}, to include only those nodes with similar climatology to the climatology of Y_i. The k^{th} nearest neighbors obtained for each station is matched with similar algorithm to the original K2. They found that it was more efficient to use a smaller set of neighbors, up to 10, as candidate parents.

3.2.2.3 The DAG Search Method

The DAG (directed acyclic graph) search algorithm is a greedy search through neighbor structures. At each step of a greedy search there is a list of all possible changes that can be made to a graph. According to Stephenson (2000), these changes can be:

- Adding an edge in any direction of the graph;
- Reversing the direction of any edge in the graph; and
- Removing an edge from the graph.

So at each step, one can choose the change which increases *log p (G)*, the second element of (3.5). Iterate until there is no change that will increase the probability. Only those changes that keep the graph as DAG are accepted.

3.3 Bayesian Networks Inference/Querying

The main task of inference is to establish the posterior probability of a set of query variables/nodes given the state of some observed event. It is to estimate the hidden nodes, given the values of the observed nodes. It is also to find the most probable configuration of a set of variables given the evidence, (Heckerman, 1999).

Inference may either be exact or approximate, (Murphy, 2002). To obtain exact inference a large dataset is required, but is computationally expensive. An approximate inference may be approximate for BNs where exact results are not strictly required.

3.3.1. Exact Inference

3.3.1.1 Pearl's Belief Propagation

Belief propagation, (Pearl & Verma, 1991), is the action of updating the beliefs in each variable when observations are given to some of the variables.

Belief propagation proceeds as follows: Let e be the set of values for all observed variables. For any variable X, e can be split into two subsets: e_X^- which represents all of the observed variables that are descendants of X, and e_X^+, which represent all of the other observed variables. The impact of the observed variables on X can be represented by the following two values:

$$\lambda(X) = P(e_X^- | X) \quad \text{and} \quad \pi(X) = P(X | e_X^+) . \tag{3.9}$$

$\lambda(X)$ and $\pi(X)$ are vectors whose elements are associated with each of the discrete values for X:

$$\lambda(X) = [\lambda(X = x1), \lambda(X = x2), \ldots, \lambda(X = x_n)], \tag{3.10}$$

$$\pi(X) = [\pi(X = x1), \pi(X = x2), \ldots, \pi(X = x_n)]. \tag{3.11}$$

The posterior probability of X given observations can be expressed:

$$P(X|e) = \alpha . \lambda(X) . \pi(X) . \tag{3.12}$$

where α is a normalizing constant.

The vector $\lambda(X)$ is computed by:

$$\lambda(X) = \prod_{c \in parents(X)} \sum_{v \in c} \lambda(v).P(v \mid X) \quad . \qquad (3.13)$$

Similarly $\pi(X)$ is computed as:

$$\pi(X) = \sum_{y \in parents(X)} P(X \mid y.\pi(y) \quad . \qquad (3.14)$$

The values of λ's, and π's are passed between nodes.

3.3.1.2 The Junction Tree Algorithm

The method involves performing belief propagation on a modified graph called a *Junction Tree*, (Stephenson, 2000). The main task is to eliminate cycles by clustering them into single cycles.

Pearl's Belief Propagation method has problems with doing inference in Directed Acyclic Graphs (DAG) when directionality is removed, due to the resulting cycles. The Junction Tree Algorithm transforms the DAG of nodes to a tree of cliques.

According to Stephenson, (2000), the junction tree algorithm proceeds as follows:

a. Moralize the graph - connect the nodes of the graph that have a common child, i.e. connect together 'unmarried' parents who share a common child, then make all edges in the graph undirected;
b. Triangulate the moralized graph - ensure that there are no childless cycles of length greater than four;
c. Construct a junction tree from the triangulated graph. Form a Maximal Spanning Tree (MST) - a selection of edges of the graph from the tree spanning every vertex;
d. Propagate the probabilities (λ and π values throughout the tree to do inference).

Figures 9a and 9b are examples of moral and triangulated graphs from Stephenson (2000).

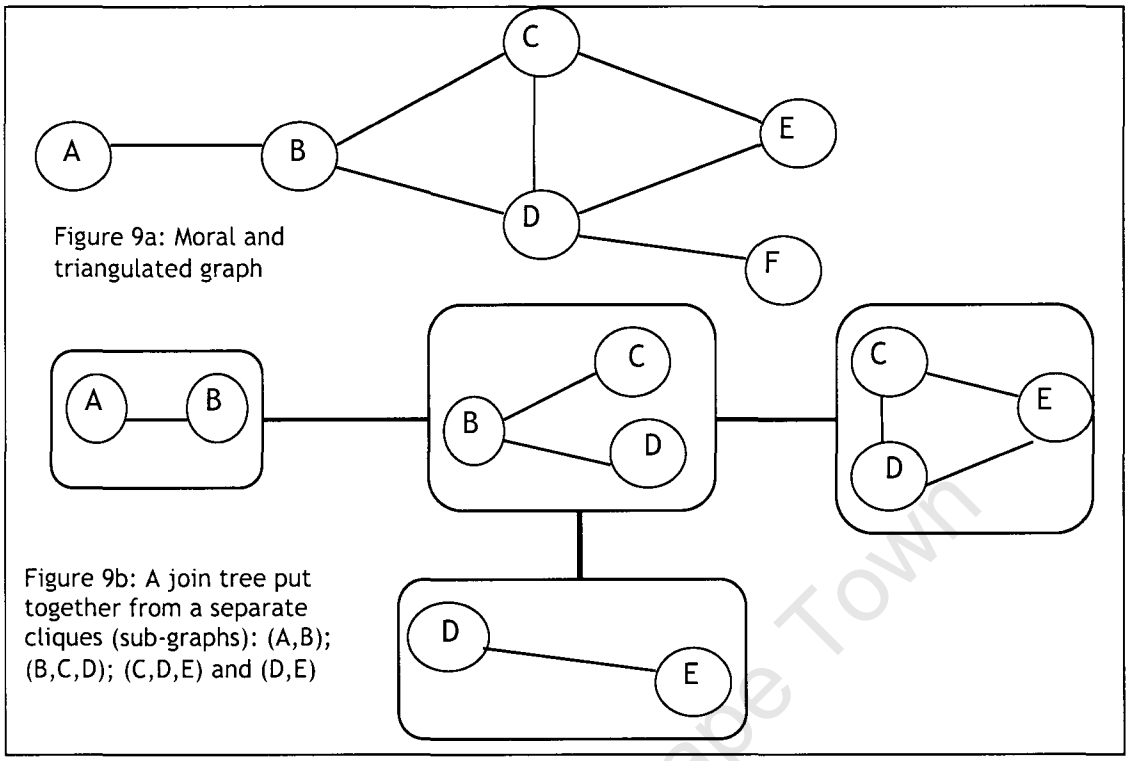


Figure 9a: Moral and triangulated graph

Figure 9b: A join tree put together from a separate cliques (sub-graphs): (A,B); (B,C,D); (C,D,E) and (D,E)

3.3.1.3 Variable Elimination

Another common exact inference method is the variable elimination. The main idea is to "push the sums in" as far as possible when marginalizing (summing) out irrelevant terms. A detailed description of this method is given by Murphy (1998). This method basically eliminates (by integration or summation) the non-observed or non-query variables one by one by distributing the sum over the product.

3.3.2 Approximate Inference

3.3.2.1 Variational Methods

Inference in some graphical models is intractable (NP-hard). Variational methods simplify the inference in graphical models by using approximation. Ghahramani (1997) explains this method in detail. The approach to approximate a probability distribution P is to define a parameterized distribution Q and vary its parameters so as to minimize the distance between Q and P.

Another approach is to transform nodes in the graph one at a time or reintroduce one node at a time in a graph. Iteratively update parameters in each node so as to minimize *cross-entropy* between the approximate and true probability distributions

(Murphy, 1998). *Cross-entropy* is an information theoretic distance between the distributions of model prediction and experimental observations.

3.4 Dynamic/Temporal Bayesian Networks

There are limitations and problems of modeling sequential data and multi-dimensional inputs with classical approaches to time series prediction, for example, with neural networks or decision trees, (Zukerman & Albretch, 2004). It is difficult to incorporate prior knowledge with these classical approaches, (Murphy, 2002).

The solution lies with state-space models. These models assume an underlying hidden state (query) that generates the observations (evidence), and that this hidden state evolves in time as a function of inputs, (Murphy, 2002). A general state-space model is the dynamic Bayesian networks (DBNs).

A dynamic Bayesian network is an extension of a Bayesian network that represents sequences of variables that are either in time-series or sequences of symbols. The model describes a system that is dynamically changing or evolving over time and enables users to monitor and predict future behavior of the system (Milhajlovic & Petkovic, 2001).

Stephenson (2000) suggested that to construct a DBN, the following needs to be defined:

- A prior network
- A transition network

Given the prior and transition networks, one can construct a DBN of any length.

DBN can be described as consisting of probability distribution function on the sequence of T hidden state variables: $X = \{x_0, \ldots\ldots, x_{T-1}\}$ and the sequence of T observable variables: $Y = \{ y_0, \ldots\ldots, y_{T-1}\}$, where T is the time boundary for the given event being investigated, (Milhajlovic & Petkovic, 2001). The system can be investigated by the following term:

$$P(X,Y) = \prod_{t=1}^{T-1} P(x_t \mid x_{t-1}) \prod_{t=0}^{T-1} P(y_t \mid x_t) P(x_0). \qquad (3.15)$$

The three sets of parameters on the right hand side of the above equation are:

- $P(x_t \mid x_{t-1})$ is *state transition* probability distribution function (pdf) that specifies time dependencies between states.
- $P(y_t \mid x_t)$ is *observation* pdf that specifies dependencies of observation nodes regarding other nodes at time slice t.
- $P(x_0)$ is *initial state* distribution that brings initial probability distribution in the beginning of the process. (Milhajlovic & Petkovic, 2001 & Petkovic, 2001)

Figures 10a - 10c adapted from Stephenson (2000), illustrate a prior network at t=0, the initial time slice, transitional for a transitional time slice and a DBN for 4 time slices. For each time slice the probabilities for each variable conditioned on other variables should be indicated. Except for the prior network for the initial time slice having its own probability distributions, the set of variables and probability definitions for transitional network are the same for each time slice.

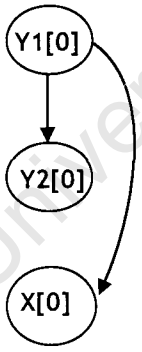


Figure 10a: Prior network at t = 0, the initial time slice.

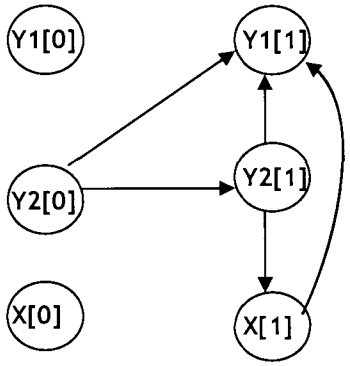


Figure 10b: Possible transition network defining a DBN for states Y1, Y2, X, incorporating the initial time slice.

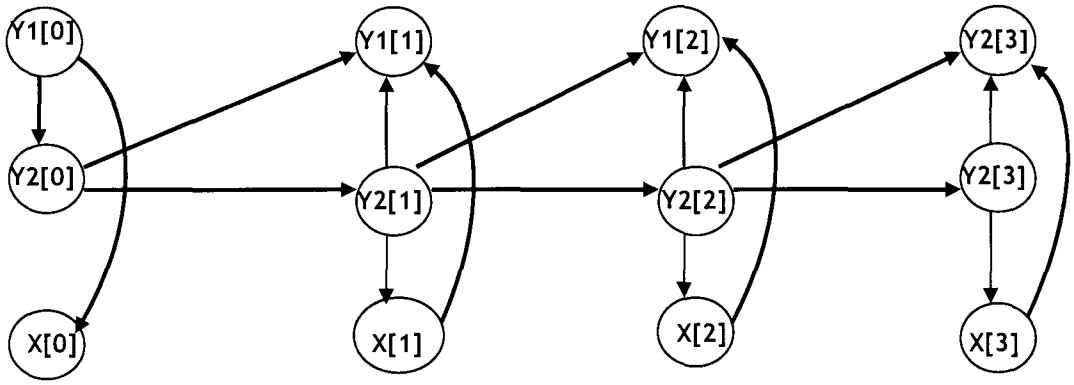


Figure 10c: A Dynamic Bayesian Network for four time slices, t=0, 1, 2, 3, derived from the prior and transition networks (figures 10a and 10b above).

3.5 Hidden Markov Models

A Markov Model is a simple model for a partially observable stochastic domain. It has, say N states, $S_1, S_2, \ldots\ldots\ldots, S_N$, with time steps t = 0, t = 1, ……,. On the t^{th} time step, the system is at one of the available states, say P_t given by

$$P_t \in \{S_1, S_2, \ldots, S_N\}. \qquad \text{(Gharamani, 1997)}$$

Between each time step, the next state is randomly chosen. The current state determines the probability distribution for the next state.

A Hidden Markov Model (HMM) is a simplified dynamic Bayesian network. In a Hidden Markov Model, the sequence of observations is modeled by assuming that each observation depends on a discrete hidden state, (Gharamani, 1997)

An example of a HMM is modeling simple states of weather and typical responses of individual human activity. Assume two common atmospheric outcomes or *states*: "Rainy" and "Sunny". These are not observed directly, they are *hidden*. Depending on the outcome of state of the weather, an individual may decide to "clean up", "drive out" or "stay in". These activities are the *observations*.

The example below from Murphy and Mian (1999) is a HMM unrolled for four time slices.

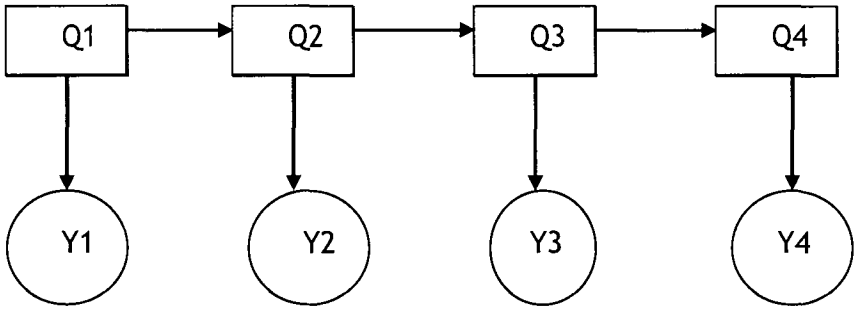


Figure 11: A Hidden Markov Model unrolled for four time slices. Squares denote discrete nodes, circles denote continuous nodes. Clear diagrams represent hidden state, while shaded ones represent observed states.

3.6. Dynamic Bayesian Network Learning

If parameters are known, the Maximum Likelihood (ML) learning algorithm is used to adjust parameters to achieve the best fit to a set of training data. A detailed discussion of the ML algorithm can be found in Grahamani (1997).

As in Bayesian networks, similar learning algorithms are applicable to DBNs (Murphy & Mian, 1999). If all parameters in a system are known, then one is able to use a DBN directly to perform inference. But this is not always so. To complete a DBN, expert knowledge in a given field is used to complement our knowledge of the problem domain, observations and experiments.

Given a dataset $D = \{Y^1, ..., Y^N\}$ of independent and identically distributed observations, each observation can be a vector or time series of vectors. Let M be a single model structure or implicit conditioning of model structure, θ be the estimated parameter. The ML can be expressed as:

$$P(D \mid \theta, M) = \prod_{i=1}^{N} P(Y^i \mid \theta, M), \qquad (3.16)$$

from (Ghahramani, 1997).

Where: P(M) is a prior probability distribution over model structures,

θ is the estimated parameter

$P(D \mid \theta, M)$ is the Maximum Likelihood

If M is dropped for notational convenience, the Maximum Likelihood parameter is obtained by maximing the log likelihood:

$$\ell(\theta) = \prod_{i=1}^{N} \log P(Y^i \mid \theta)$$

If the observational vector includes all variables in the BN, then each term in the log likelihood factors as:

$$\log P(Y^i \mid \theta) = \log \prod_j P(Y_j^i \mid Y^i{}_{P_a(j)}, \theta_j)$$

$$= \sum_j \log P(Y_j^i \mid Y^i{}_{P_a(j)}, \theta_j), \qquad (3.17)$$

Where 'j' indexes over the nodes in the network, $P_a(j)$ is the set of parents of j, and Θ_j are the parameters that define the conditional probability of Y_j given its parents.

Maximum Likelihood algorithm may not be used in the case of hidden states in DBN (Milhajlovic & Petkovic, 2001). These are values that cannot be determined in every time slice. In such cases, the Expectation Maximization (EM) learning algorithm is used.

Let u_i and o_t represent unobserved and observed variables respectively for time slice t. The probability distribution can be written:

$$P(o_1,\ldots,o_T,u_1,\ldots,u_T) = P(u_1)P(o_1 \mid u_1)\prod_{i=2}^{T} P(u_t \mid u_{t-1})P(o_t \mid u_t). \qquad (3.18)$$

Where: u_i and o_t are unobserved and observed variables,

If we let θ denote model parameter and θ_i^* the optimized model parameters, then the EM learning algorithm are as follows, (Milhajlovic & Petkovic, 2001):

EM algorithm:
Get initial guess of θ_0;
do
infer hidden variables $\hat{u}_t \; o_1,\ldots,o_T$, using model θ_k;
$\theta^*_{k+1} = \arg_\theta \max P_r(o_1,\ldots,o_T,\hat{u}_1,\ldots,\hat{u}_T \mid \theta)$;
until (convergence).

3.6.1 *Forward and Backward propagation algorithm*

The forward-backward algorithm is belief propagation applied to Bayesian network, (Ghahramani, 1997). In a DBN only a subset of states can be observed at each time slice, and it is necessary to calculate unknown states.

From the figure 13 below, given observations in every time slice y_i, we have to estimate values of hidden nodes x_i, where i receive values from 0 to T-1.

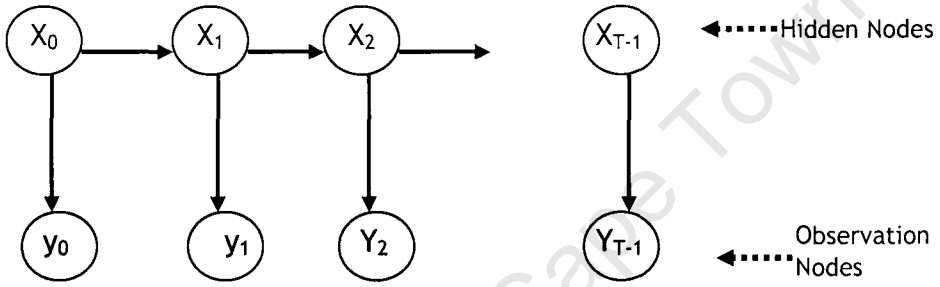


Figure 12: A DBN structure, taken from Ghahramani (1997) and Milhajlovic & Petkovic (2001)

The forward pass recursively computes α_t defined as the joint probability distribution of Y_t, and the sequence of observations gathered from t to t+1, the forward probability distribution is

$$\alpha_t(x_t) = P(Y^t_o, x_t). \tag{3.19}$$

Given the network structure shown above, from equation (3.19), (Milhajlovic & Petkovic, 2001) conclude that

$$\alpha_{t+1}(x_{t+1}) = P(y_{t+1} \mid x_{t+1}) \sum_{x_t} P(x_{t+1} \mid x_t)\alpha_t(x_t) \ . \tag{3.20}$$

The backward pass computes the conditional probability $\beta_t(x_t)$ of the observations from time t+1 to last observation at time t-1, defined

$$\beta_t(x_t) = Pr(Y_{t+1}{}^{T-1} \mid x_t). \tag{3.21}$$

From the definition, we can similarly arrive at the relation:

$$\beta_{t-1}(x_{t+1}) = \sum_{x_t} \beta_t(x_t) P_r(x_t \mid x_{t-1}) P_r(y_t \mid x_t). \quad (3.22)$$

The final value being $\beta_T(x_{t-1}) = 1.$

3.6.2 Pruning

Apart from learning parameters as in BNs, we also learn structure in DBNs. It may be necessary to do *pruning* in a DBN structure. It is a complex process and involves distinguishing which nodes are important for inference in a DBN structure and which are not (Milhajlovic & Petkovic 2001).

The process may consist of one of the actions: nodes that are not important may be deleted or removed from the network. The states may also be deleted from a particular node or connections between two nodes are removed.

The process is carried out while maintaining accuracy, i.e. no loss of information. If a DBN sensor failed, then there is a chance that data was incorrect (Milhajlovic & Petkovic 2001).

3.7 DBN Inference

The main task of inference in a DBN is to estimate the probability distribution function (pdf) of unknown states given some known observations and initial probability distribution.

In the DBN, only a subset of the states can be observed at each time slice, all the other unknown states in the network have to be determined or calculated. The procedure of calculation is called *inference*.

If we have a finite set of consecutive observations $Y^{T-1} = \{y_0, \ldots\ldots, y_{T-1}\}$, and a set of corresponding hidden variables, $X^{T-1} = \{x_1, x_2, \ldots\ldots, x_{T-1}\}$, the task of inference is to find a conditional pdf, $P(X^{T-1} \mid Y^{T-1})$.

Depending on specific structure of a DBN, it may be more appropriate to estimate the pdf's statistics - mean, variance or covariance, instead of conditional pdfs.

According to Murphy, (1998), the general problem in inference for dynamic Bayesian networks is to compute $P(X(i, t) \mid Y(:, t1:t2))$, where

$X(i, t)$ represents the i^{ths} hidden variable at time t;

and $Y(:, t1:t2)$ represents all evidence at times t1 and t2.

Murphy, (1998), also suggested that we may also want to compute the joint probability distribution of variables over one or more time slices.

The figure below is a summary of the main kinds of inference for DBNs: *filtering, prediction, fixed-lag smoothing and fixed interval smoothing.*

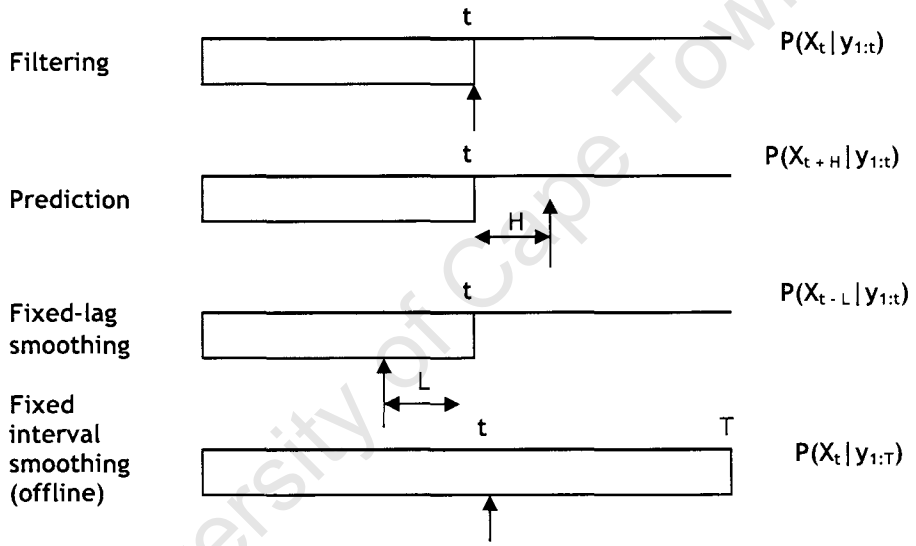


Figure 13, after Murphy (2002): Summary of the main kinds of inference for DBNs. The shaded region is the interval for which we have data. The arrow represents the time step at which we want to perform inference. t is the current time, and T is the sequence length. H is a prediction horizon and L is a time lag.

From the above figure, we summarize inference problems that we might be interested in:

(a) *Filtering*: computing $P(X_t \mid y_{1:t})$, i.e., monitoring (tracking) the state over time. Filtering method involves estimating belief state i.e the posterior distribution over the current state, given all evidence to date, (Russel & Norvig, 2002). We compute $P(X_t \mid y_{1:t})$ given that $P(X_{t-1} \mid y_{1:t-1})$ evidence arrives on time beginning at time t=1.

Using Bayes' rule, (Murphy, 2002):

$$P(X_t \mid y_{1:t}) \alpha P(y_t \mid X_t, y_{1:t-1}) P(X_t \mid y_{1:t-1}) = P(y_t \mid X_t)\left[\sum x_{t-1} P(X_t \mid x_{t-1}) P(x_t \mid y_{1:t-1})\right]. \quad (3.23)$$

where the constant of proportionality is $1/ct = 1/P(y_t \mid y_{1:t-1})$.

Following the example of 'umbrella/rain world' advanced by Russel & Norvig, (2002), in weather and climate assessment, this would mean computing the probability of onset of rains in the month of November, given all the observations of other weather parameters for October.

(b) *Prediction*: Apart from estimating the current or past state, it is necessary to predict the future. This is done by computing $P(X_{t+H} \mid y_{1:t})$ where $H > 0$ is how far ahead in future we would like to predict. Future observations can be predicted by marginalizing out X_{t+H}, (Murphy, 2002):

$$P(Y_{t+H} = y \mid y_{1:t}) = \Sigma_x P(Yt+H = y \mid Xt+H = x) P(Xt+H = x \mid y1:t) \quad (3.24)$$

Prediction is a two step process: prediction (to compute $P(X_t \mid y_{1:t})$) and updating (to compute $P(X_t \mid y_{1:t})$).

Using the 'umbrella/rain world' by Russel & Norvig, (2002), in our weather and climate assessment, this may mean computing the probability of a normal onset of rains, 3 months before the onset, given the present observations in East Pacific SOI, Indian Ocean SST and SLP.

(c) *Fixed-lag smoothing*: We may be interested to estimate the state of the past given the evidence up to the present. We compute $P(X_{t-L} \mid y_{1:t})$, where $L < 0$ is the lag. In the weather scenario, we could compute the probability that there was a late onset of rains the immediate past year 2007, given observations in January, 2008.

(d) *Fixed-interval smoothing*: It corresponds to computing $P(Xt \mid y_{1:T})$ for all $1 \leq t \leq T$. This is used as a subroutine for offline training. In the weather and Russel and Norvig, (2003) suggest that this can be done in two parts - computing the evidence up to time t, and computing the evidence from t+1 to T.

In the weather scenario, we could compute the probability of, say, a false onset the previous two months, given the probability of a normal onset of rains and other observed parameters in the current month.

3.8 Emergent Situation Awareness Technology

A technology used in this project is the Emergent Situation Awareness (ESA). It was developed by Osunmakinde and Potgieter (2009). The technology is an innovative Dynamic Bayesian network (DBN) technology, which truly evolves temporal models and reveals what is currently happening over time in a domain of interest. One of its powerful features is its evolvement from multivariate time series (MTS) data in the absence of domain experts. The ESA advances the algorithms of ordinary Bayesian networks to evolve dynamically as it changes its network and the probabilistic distributions with time. It has three notable components, which include learning algorithms, a probabilistic reasoner and trend analysis.

The ESA learning uses genetic algorithm to emerge temporal Bayesian networks called frames over the time steps from the MTS environments. The probabilistic reasoner is the Bayesian inference, which does necessary forward and backward propagations through the links of the frames and generates probable results. The trend analysis is an interface that generates n-dimensional transition matrices of knowledge, where n corresponds to the pieces of knowledge to be revealed e. g. a transition matrix of target probabilities, a transition matrix of target parameter values, etc. For users' interaction and simple decision making, the ESA requests answers to the following questions: (i) What is happening? (ii) Why is it happening? (iii) What will happen next? (iv)What can I do about it?

The ESA was used for experiments to prepare an academic paper from this dissertation for publications. The paper describes how to understand the variability of rainfall in southern Africa (Cheruiyot, Potgieter & Osunmakinde, 2008). (See Appendix 2 for reference to this paper).

For assessment of rainfall and onset of rainfall, it was best to study and understand it in recurrent time, i.e. month and not year. The month parameter is recurrent since it is repeated in real life, unlike year, say 1999, which will never be repeated in real life. January comes every year, but 1999 will never recur. The ESA technology was to focus on understanding the variability of the onset of rainfall.

3.9 Conclusion

In this chapter we have explored various aspects of Bayesian networks: learning/training; BN structures and Inference. We have also explored theoretical background of Hidden Markov Models (HMM) and dynamic Bayesian networks (DBN). We have seen that DBN can be used to model systems that are dynamically changing or evolving over time and can enable users to monitor and further predict the behavior of the system.

Chapter 4: Data Analysis and Methodology

4.0 Introduction

In this chapter we describe sources of data used in this work, the domains where sea surface temperatures and sea level pressures were extracted in the global oceans and how the data was organized from these sources.

We also describe how rainfall data was analyzed to determine the Start of Season (SoS)/dates of onset of rains. We shall also describe the method of Principal Component Analysis that was used to obtain homogeneous rainfall zones in Botswana for the months of October, November and December, (OND).

It was also necessary to include in this chapter a description of data organization for analysis with Emergent Situation Awareness (ESA) technology for dynamic Bayesian networks.

4.1 Organization of Data

The data used in this study were daily, pentad and monthly rainfall in Botswana for the years 1971 - 2004. It was obtained from Botswana Meteorological Services. Daily rainfall data available is arranged such that day 1 is July 1st and the last day (day 366) is June 30th the following year. Stations that had many gaps in rainfall data or had more than 10% of monthly data missing were discarded.

El-Niño/Southern Oscillation (SOI) data were obtained from NOOAs National Center for Environmental Prediction/Climate Prediction Center (NCEP/CPC) and European Center for Medium-range Weather Forecasting (ECMWF). The data is available from websites: http://www.cpc.noaa.gov/data/indices; http://climexp.knmi.nl/; http://www.ecmwf.int/research/era.

The other atmospheric data we used in this project are the monthly SLPs, SST anomalies are at a 2^0 X 2^0 spatial resolution for the period January 1970 to 30 August 2002. They were extracted from: ECMWF Re-Analysis (ERA-40) (1957-2000) period and NCEP/NCAR (1948-present). Anomalies for each of the fields were based on climatology for the period November 1979 to July 1999, (Kalnay, 1999). The 500

hPa geopotential height anomalies and 700 hPa zonal and meridional windspeeds (u and v respectively), are gridded to the same spatial domain as described for the SST dataset.

The SLPs were extracted for the following selected domains:

- South Indian Ocean: 25^0S - 35^0S, 37^0E - 50^0E;
- Central Indian Ocean: 10^0S - 20^0S, 52^0S - 60^0S;
- Atlantic Ocean: 15^0S - 30^0S, 10^0W - 10^0E.

The SSTs were extracted for the following selected domains, (based on results of table 4):

- Central Indian Ocean: 5^0S - 15^0S, 60^0E - 85^0E;
- Atlantic Ocean: 15^0S - 30^0S, 10^0W - 10^0E.

These are regions where semi-permanent Sub-Tropical High Pressure synoptic systems are found, the Mascarene High and the St-Helena High respectively. For the two SST domains, the regions were also determined by correlating Botswana rainfall with Global SSTs. The correlation analysis methodology and result is described in section 3.4.2.

Table 3 is a list of 29 stations that were selected for use in this study.

Table3: List of rainfall stations used

STN #	STATION	STATION FULL NAME	Lat °S Deg	min	Long °E Deg	min
002	BAINES	Baines Drift Police Stn.	22	29	28	44
006	BOBO	Bobonong Police Stn.	21	58	28	25
024	DIBET	Dibete Police Station	23	50	26	25
030	DUKWE	Dukwe Police Station	20	32	26	34
041	GD_HOPE	Good Hope Agr. Res. Stn.	25	28	25	26
043	GUMAR	Gumare Agric. Stn.	19	22	22	10
044	GWETA	Gweta Police Stn.	20	12	25	14
065	KAVIMB	Kavimba Police Stn.	18	5	24	35
069	KGALE	Kgale (St. Josephs Sch)	24	42	25	52
094	LETN	Letlhakeng police Station	24	04	25	02
096	LOBAT	Lobatse Police Stn.	25	15	25	39

103	MACHNG	Machaneng Police Stn.	23	11	27	30
121	MART	Martin's Drift Police St	22	59	27	56
136	MOCHDI	Mochudi Police Stn.	24	23	26	8
147	MOLEPS	Molepolole Police Stn.	24	23	25	30
164	NATA	Nata Police Station	20	13	26	10
166	NCOJANE	Ncojane Police Station	23	9	20	17
177	OLF_DRFT	Olifantsdrift Police Stn	24	12	26	41
179	ORAPA	Orapa Power Station	21	19	25	23
180	PMOL	Phitshanemolopo Police	25	45	25	6
182	PALAPY	Palapye Teba	22	34	27	9
183	PANDA	Pandamatenga Police Stn.	18	33	25	38
186	PELOTS	Pelotshetlha Primary Sch	25	11	25	22
195	RAKOPS	Rakops Police Station	21	3	24	24
198	RAMATL	Ramatlabama Police Stn.	25	39	25	34
199	RAMOTSW	Ramotswa Station	24	53	25	52
207	SEBINA	Sebina Store	20	51	27	13
211	SEHITW	Sehitwa Police Station	20	20	22	24
220	SEROW	Serowe Police Station	22	23	26	42
223	SHAK	Shakawe Met Station	18	22	21	50
241	TONOTA	Tonota Police Station	21	27	27	28
244	TSHAB	Tsabong Airport	26	0	22	24
251	TSHA	Tshane Met Station	24	1	21	53
252	TSHETS	Tshesebe Police Station	20	41	27	35
256	TUTUME	Tutume Police Station	20	30	27	1

4.2 Wet and Dry Years

To determine major wet and dry years, we used the following method, from Okoola (1999) and Philippon (2002):

$$X_t = \frac{1}{m}\sum_{j=1}^{m}\frac{100X_{tj}}{\overline{X_j}} \qquad (4.1)$$

Where m is the total number of stations used, X_t is the time-dependent rainfall index as a % of mean and averaged over the whole country.

Rainfall data was normalized to ensure that the mean is zero and standard deviation was unit using the following equation:

$$x = \frac{X - \overline{X_c}}{\sigma_x}$$, x is the standardized seasonal anomaly index;

where X is the observed value; σ_x the standard deviation; and $\overline{X_c}$ is the mean.

"Rainfall performance" was deduced from rainfall index, X_t, The ***wet*** years had large values of the index (X_t >125%) while ***dry*** years had low values (X_t <75%). ***Normal*** years had values of this index had values between 75% and 125%. Abnormally low values are referred to as ***very dry***, while abnormally high values were referred as ***very wet***.

Figure 14 below shows resulting standardized rainfall anomalies[2] and time series of Botswana rainfall.

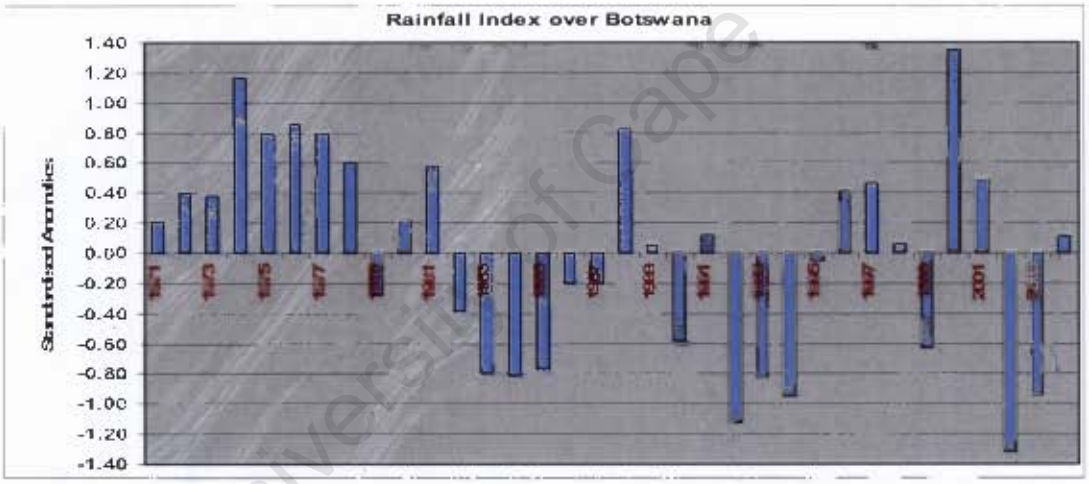


Figure 14: Standardized anomalies of annual rainfall over Botswana.

4.3 Determination of Start of Season

Statistical analysis was done using daily rainfall data for several stations. For each selected station, for the years 1971 to 2004, daily rainfall data was arranged in tables from day 1 (1st July of each year) to day 366 (June 30th of the following year). Appropriate codes were used to represent missing data and February 29th for non-leap years. Since Botswana has a semi-arid climate with large annual rainfall variability, subjective criteria was used to determine the *start of season* for each station.

[2]Anomalies in this context means departures from the normal

The *Start of Season (SoS)* used for this study was based on the following definition: ***the date after 1st October when cumulative rainfall exceeded 20 mm, and that no dry spell exceeding 20 days occurs in the next 30 days***. This definition was adopted and modified from that used by Famine Early Warning System (FEWS) for Southern African countries, (FEWS: http://edcintl.cr.usgs.gov/fews/wrsi.html).

The following additional criteria were also used in the analysis:

- For each station, the first day (day 1) on the worksheet was July 1st, while the "start of the season", or the earliest day, *Start of Season* was taken as October, 1st, that is day 93.
- A rainy day is the day when there was recorded at least 0.85 mm of rain.

Statistical summary was done for each year for each station to determine the mean onset date, standard deviation, minimum and maximum onset dates. Graphs for *Start of Season* for each station were plotted. Types of onset were determined for each year and categorized as *False (0), Early (1), Normal (2), Late (3) and Failed (4) Onsets.*

Figure 15 is a time series plot of rainfall onset and categories of onsets for Botswana. The plot can be summarized and interpreted thus:

- *Mean Start of Season (SoS) = 25th November (Day 148).*
- *Normal Onset = Mean SoS* +/- 10 days (Days 138-158)
- *Early Onset = Earliest normal onset date - 20 days* (Days 138 - 118)
- *Late Onset = Latest normal onset date + 20 days.* (Days 158 - 178)
- *Failed Onset= Days later than Day 178 (25th Dec)*
- *False Onset = Days earlier than Day 118 (26th Oct).*

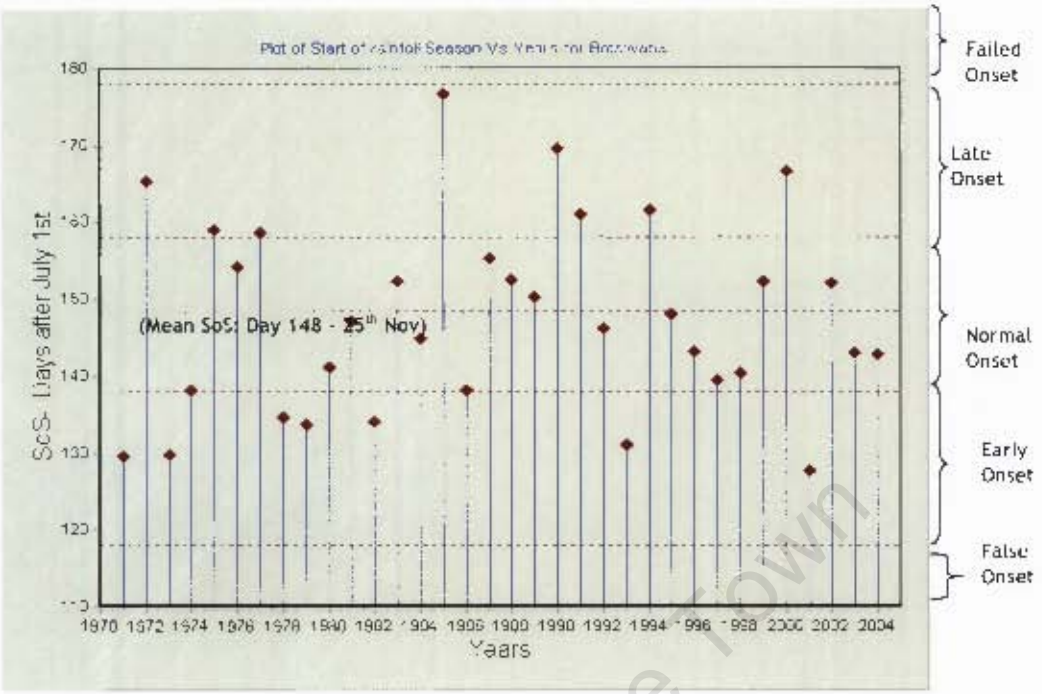


Figure 15: Time series plot of rainfall onset in Botswana

An example Onsets/Start-of-Season for Serowe rainfall station (Central Botswana) is illustrated in the figure below using the above criteria. As in the previous plot (figure 15) we can interpret and summarize the onset of rains in Serowe thus:

- *Mean Start of Season (SoS) = 14th November (Day 137).*
- *Mean Onset = Mean SoS +/-* 10 days (Days 127-147)
- *Early Onset = Earliest normal onset date - 20 days* (Days 127 - 107)
- *Late Onset = Latest normal onset date + 20 days.* (Days 147 - 167)
- *Failed Onset= Days later than Day 167 (14th December)*
- *False Onset = Days earlier than Day 107 (15th Oct).*

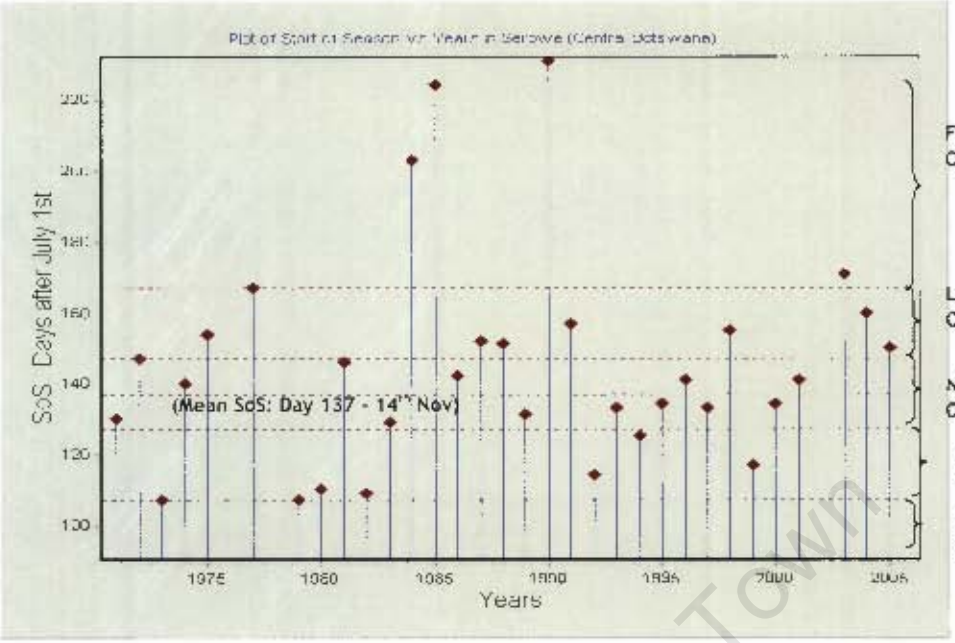


Figure 16: Time series of Start of season in Serowe (Central Botswana)

The table below gives a summary of rainfall indices and national rainfall onsets.

Table 4: Summary of rainfall indices and national rainfall onsets

YEAR	RAINFALL INDEX, N_i	SEASONAL ANOMALY INDEX, x	RAINFALL PERFORMANCE	NATIONAL RAIN ONSET	COMMENTS ON ONSET
1971	107.85	0.21	NORMAL	EARLY	
1972	114.67	0.40	NORMAL	NORMAL	Normal rains, Late Onset in the North
1973	113.85	0.38	NORMAL	EARLY	
1974	143.50	1.16	VERY WET	NORMAL	
1975	130.11	0.79	WET	NORMAL	Wet years and Normal Onset
1976	131.98	0.86	WET	NORMAL	
1977	129.13	0.80	WET	LATE	Wet year, late onset
1978	122.94	0.61	NORMAL	EARLY	Normal rainfall, Early Onset
1979	90.40	-0.28	NORMAL	EARLY	
1980	106.60	0.22	NORMAL	NORMAL	Normal rains, normal onset, early in Central Region
1981	121.06	0.57	NORMAL	NORMAL	
1982	85.83	-0.83	NORMAL	EARLY	
1983	70.93	-0.79	DRY	NORMAL	Dry Years, mixed onsets. 1984-85 late onset in Central Region
1984	70.10	-0.81	DRY	NORMAL	
1985	71.28	-0.77	DRY	LATE	
1986	93.22	-0.20	NORMAL	NORMAL	Normal rains, normal onset
1987	92.70	-0.21	NORMAL	NORMAL	
1988	130.09	0.82	WET	NORMAL	
1989	102.02	0.05	NORMAL	NORMAL	

1990	78.66	-0.58	NORMAL	LATE	Normal Rains, Late Onset
1991	104.06	0.12	NORMAL	LATE	
1992	58.24	-1.12	VRY DRY	NORMAL	Very Dry and Normal onset. False in Kalahari
1993	69.41	-0.82	DRY	EARLY	Dry and Early Onset
1994	64.72	-0.94	DRY	LATE	Very Dry and Late onset
1995	98.39	-0.05	NORMAL	NORMAL	
1996	115.11	0.41	NORMAL	NORMAL	Normal Rains, and Normal Onset
1997	117.01	0.46	NORMAL	NORMAL	
1998	102.57	0.06	NORMAL	NORMAL	
1999	76.33	-0.63	NORMAL	NORMAL	
2000	149.89	1.35	WET	LATE	Very Wet and late onset
2001	118.92	0.49	NORMAL	EARLY	Normal rains, early onset. Failed Onset Kalahari
2002	51.63	-1.31	DRY	NORMAL	Very dry, normal onset in North and very late/failed onset in Central Region
2003	66.64	-0.94	DRY	NORMAL	
2004	104.12	0.11	NORMAL	NORMAL	Normal year

4.4 Determination of Homogeneous zones

4.4.1 Principal Component Analysis

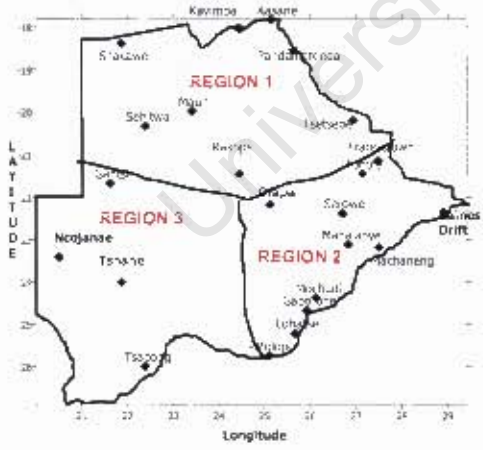
The method of Principal Component Analysis (PCA) was used to delineate rainfall homogeneous zones in Botswana. Seasonal rainfall data in spatial "s-mode" was used in each station and for all years (1971 - 2004). The data was normalized to ensure that the mean is zero and the Standard Deviation was unity. This topic is explained further in Appendix 1.

Using statistical software, the standardized data was then used to determine significant principal components. The significant components were rotated using varimax method. The varimax criterion maximizes the variance of the squared correlation coefficients (squared loadings) between each rotated principal component, (Bazira & Ogallo, 1999). The spatial patterns of the dominant PCA modes were then used to group together stations with similar characteristics into homogeneous zones.

Table 5: Results from Principal Component Analysis

	Latent Roots (Eigenvalues)	% Total Variance Explained	Cumulative Variance
1	11.03	38.150	38.150
2	3.977	13.715	51.865
3	1.995	6.878	58.743
4	1.658	5.718	64.461
5	1.571	5.417	69.878
6	1.182	4.077	73.955
7	1.017	3.505	77.460

The results of the PCA indicated that the seasonal rainfall over Botswana could be represented by seven (7) components which explained 77.5% of the total variance. The spatial patterns of the significant loadings highlighted the different climatological zones in the country. However some of the representative regions, for example areas around Rakops and Baines Drift in Limpopo valley, were small or localized. Taking into consideration the climatology and rainfall patterns in the country, I decided to delineate the country into three major homogeneous zones, shown in figure 17.



REGION1: NORTHERN AREA:
Stations: Kavimba, Pandamatenga, Sehitwa and Tsetsebe

REGION 2: CENTRAL & SE:
Stations: Orapa, Olifants_Drift, Lobatse, Lethlakeng, Kgale, Good-Hope, Baines-Drift and Serowe

REGION 3: WEST & KALAGADI:
Stations: Tshane and Tsabong

Figure 17: Rainfall homogeneous Zones for October, November and December (OND) in Botswana

Figure 18 shows October, November and December (OND) rainfall in Botswana.

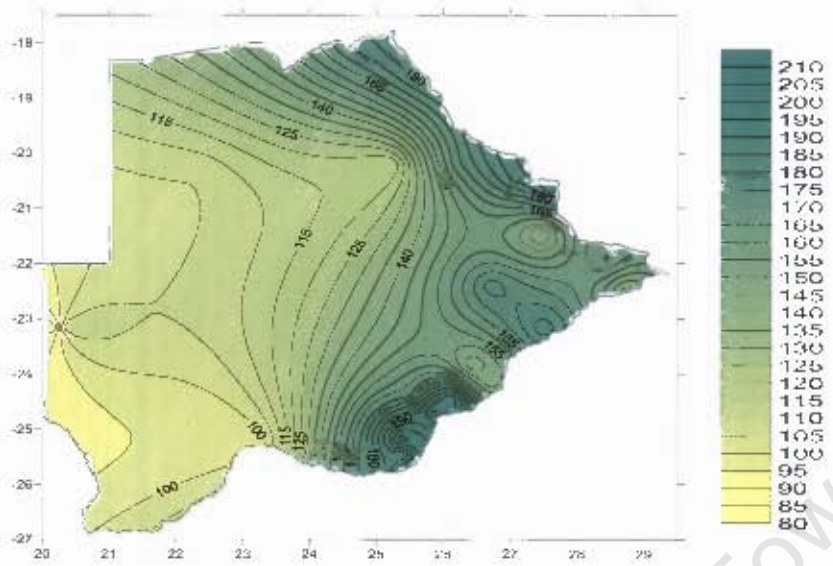


Fig 18: October, November and December (OND) Botswana rainfall

4.4.2 Correlation Analysis

Using *WinGrads* and *FORTRAN* software, the significant sea surface temperature (SST) modes in the Indian, Atlantic and Pacific Oceans for July, August and September (JAS) were correlated with October, November and December (OND) rainfall over the representative stations in Botswana. Rainfall in representative stations was correlated with three months lead time for SSTs, (T-3) in global oceans. Changes in atmospheric circulations accompanying anomalous warming, such as El-Niño, induce changes in cloud cover and evaporation, and in turn increase the net heat flux which warms the oceans, (Klein, 1999). This is the rationale for correlating OND Botswana rainfall with JAS Sea Surface Temperatures in the global oceans. Whenever there is anomalous warming or cooling in the sea surface, for example in the East Pacific during an El-Niño year, it would take approximately 3-6 months, for corresponding changes in atmospheric circulations in Tropical North Atlantic and Indian Oceans (Klein, 1999).

The significant SST modes (Pacific, Atlantic and Indian Oceans) of JAS were correlated with OND rainfall over each of the three regions in Botswana. The following are some of the results obtained:

a. Pacific Ocean

- Regions: 6^0N -20^0N, 128-146^0W and 23^0S-28^0S, 159°W-171^0W, were negatively correlated with Region 2 (Central and South East);
- Region 13^0S -21^0S, 78-90^0W correlated negatively with Region 3 (Western and Kalagadi area);
- Region 24^0S-29^0S, 94°W-98^0W correlated positively with Region 1.

b. Atlantic Ocean

- Region: 0^0N -8^0N, 39^0W-48^0W correlated positively with Region 3;
- Region 45^0S -50^0S, 50^0W-56^0W correlated negatively with Region 1.

c. Indian Ocean

- Regions: 7^0S-14^0S, 56°E-64^0E and 34^0S-38^0S, 34°E-46^0E, were correlated negatively with Region 2 (Central and South Eastern) areas.

Table 6: Some significant cross correlation coefficients (>|0.4|). between Regions 1, 2 and 3 in Botswana and global oceans (e.g. Central or East Pacific, Indian Ocean, and South Atlantic) for months July, August and September, (JAS).

	Region 1	Region 2	Region 3
July		*REG2*	
		REG2 1	
		E-PCF -0.403314	
		E-PCF -0.515939	
August		*REG2*	
		REG2 1	
		E-PCF -0.596879	
Sept	*REG1*	*REG2*	*REG3*
	REG1 1	*REG2* 1	REG3 1
	WST-PCF 0.468841831	S-PCFC 0.515594	E-PCF -0.483085
	ST-ATL -0.50542433	CTR-PCFC -0.533683	ST-PCF 0.543686
		IND/OCN -0.502186	IND/OCN -0.479394
		EST-IND/OC 0.576934	NRT-IND/OCN 0.533486

Legend: E-PCF -East Pacific; WST-PCF- West Pacific ; CTR-PCF- Central Pacific; IND/OC-Indian Ocean; EST-IND/OC- Eastern Indian Ocean; NRT-IND/OC- Northern Indian Ocean.

Table 7: Summary of rainfall index and Regional rainfall onsets (indices explained below), (From earlier result of PCA analysis, the country was subdivided to 3 regions:
REGION1: NORTHERN AREA: Stations #s: 65, 183, 211, and 252
REGION 2: CENTRAL & SE: Stations #s: 2, 41, 69, 94, 96, 177, 179, 220
REGION 3: WEST & KALAGADI: Stations #s: 244, 251)

YEAR	RAINFALL INDEX, X_t	RAINFALL PERFORMANCE	ONSET REGION1	ONSET REGION2	ONSET REGION3	NATIONAL RAIN ONSET	
1971	107.85	NORMAL	2	1	1	1	EARLY
1972	114.67	NORMAL	3	2	2	2	NORMAL
1973	113.85	NORMAL	2	1	1	1	EARLY
1974	143.50	VERY WET	2	2	2	2	NORMAL
1975	130.11	WET	2	2	2	2	NORMAL
1976	131.98	WET	2	0	2	2	NORMAL
1977	129.13	WET	3	3	1	3	LATE
1978	122.94	NORMAL	0	0	999	0	FALSE
1979	90.40	NORMAL	1	2	999	1	EARLY
1980	106.60	NORMAL	2	2	2	2	NORMAL
1981	121.06	NORMAL	2	2	1	2	NORMAL
1982	85.83	NORMAL	2	3	3	3	LATE
1983	70.93	DRY	2	2	3	2	NORMAL
1984	70.10	DRY	1	3	2	2	NORMAL
1985	71.28	DRY	999	4	1	3	LATE
1986	93.22	NORMAL	1	2	1	2	NORMAL
1987	92.70	NORMAL	2	2	3	2	NORMAL
1988	130.09	WET	3	3	3	3	LATE
1989	102.02	NORMAL	3	2	2	2	NORMAL
1990	78.66	NORMAL	2	4	3	3	LATE
1991	104.06	NORMAL	2	3	3	3	LATE
1992	58.24	VRY DRY	999	2	0	2	NORMAL
1993	69.41	DRY	2	2	3	2	NORMAL
1994	64.72	DRY	3	3	0	3	LATE
1995	98.39	NORMAL	2	2	0	2	NORMAL
1996	115.11	NORMAL	2	2	2	2	NORMAL
1997	117.01	NORMAL	2	2	2	2	NORMAL
1998	102.57	NORMAL	2	2	0	2	NORMAL
1999	76.33	NORMAL	2	2	2	2	NORMAL
2000	149.89	WET	2	2	2	2	NORMAL
2001	118.92	NORMAL	2	2	4	2	NORMAL
2002	51.63	DRY	2	2	3	2	NORMAL
2003	66.64	DRY	2	4	2	2	NORMAL
2004	104.12	NORMAL	2	2	999	2	NORMAL

ONSETS		Index
False Onset	-	0
Early Onset	-	1
Normal Onset	-	2
Late Onset	-	3
Failed Onset	-	4
Undetermined/Missing data	-	999

4.5 Organization of Data for analysis with Emergent Situation Awareness (ESA) for dynamic Bayesian networks

Daily rainfall data from 14 selected stations were organized from 1st July (Day 1) to 30th June (Day 366) of the following year for the years July 1971 to Sept 2002. Incomplete or missing data was flagged with *9999*. For leap year *(Feb29th) or Day 244*, we used 9988.

Data organization **with corresponding parameters**:

a. *Station#*[3], Year, Month
b. *Mthly_RR* - Monthly Rainfall (in mm)
c. Ind_SST_Anom -Indian Ocean Sea Surface Temperature Anomalies (^{0}C)
d. Atl_SST_Anom - Atlantic Ocean Sea Surface Temperature Anomalies (^{0}C)
e. SOI - Southern Oscillation Index (Indicator for El-Niño/La-Niña) - (Warming/cooling episodes in Eastern Pacific)
f. (T- 3)SOI - three months lag in SOI
g. Atl_SLP_Anom - Atlantic Ocean Sea Level Pressure Anomalies (HPa/mb)
h. SInd_SLP_Anom - Southern Indian Ocean (Mascarene area) Sea Level Pressure Anomalies (HPa/mb). Domain :34^{0}S-38^{0}S, 34^{0}E-46^{0}E
i. CInd_SLP_Anom - Centra Indian Ocean (North of Malagasy area) Sea Level Pressure Anomalies (HPa/mb). Domain: 7^{0}S-14^{0}S, 56^{0}E-64^{0}E
j. 500HPa_Anoms - 500 hPa hectopascal level anomalies (in meters)
k. 700HPa_U - 700 HPa level *zonal* winds (meters per second)
l. 700HPa_V - 700 HPa level *meridional* winds (meters per second).
m. Onset[4] - Onset of rains in Botswana or "National Onset".

[3] *Station#- Station Number*
[4] In preparation for analysis of data using ESA for DBN, we have included onset type in the column for each month variable (January to December). It should be noted here that in reality, there is only one onset of rains in a year that occurs during summer rainy months (October - March) in Southern Africa and can be a *false, early, normal, late or failed* onset for a given year.
For organization of data, national onset type (from Table 7) for any particular year is entered from ***July of that year to June*** the following year.

Sample data based on this organization is included in Appendix 4.
Table 7 above (an extension of table 4) is a summary of rainfall indices and onsets of rains in the three homogeneous zones and national onset.

4.6 Emergent Situation Awareness (ESA) Technology

For assessment of rainfall and onset of rainfall, it was best to study and understand it in recurrent time, i.e. *month* and not *year*. The month parameter is recurrent since it is repeated in real life, unlike year, say 1999, which will never be repeated in real life. January comes every year, but 1999 will never recur. The ESA technology was to focus on understanding the variability of the onset of rainfall.

We consider normal onsets of rainfall in all the years for January to December months given anomalies in Sea level Pressure values in Southern Indian Ocean (South of Mauritius - Mascarene High region) and Central Indian Ocean (East of Madagascar). These oceanic regions are crucial moisture source areas for Southern Africa (Washington and Preston, 2006).

The graph below (figure 20a) shows the outcome from ESA and a User's Interaction. It shows probability reasoning over time. It shows probability/degrees of beliefs for normal onset of rainfall given that sea level pressure (SLP) anomaly in south Indian Ocean (34^0S-38^0S, 34^oE-46^0E) is -1.082 hPa and SLP anomaly in Central Indian Ocean (7^0S-14^0S, 56^oE-64^0E) lies between -0.129 hPa and +2.47 hPa. Figure 20b shows emergent situation of onset for positive anomalies in South Indian Ocean; while figure 20c shows emergent situation for normal onset for a very wet year, 2000.

The figures will be used to answer the key questions: *What is happening? Why is it happening? What will happen next?* And *what can I do about it?*

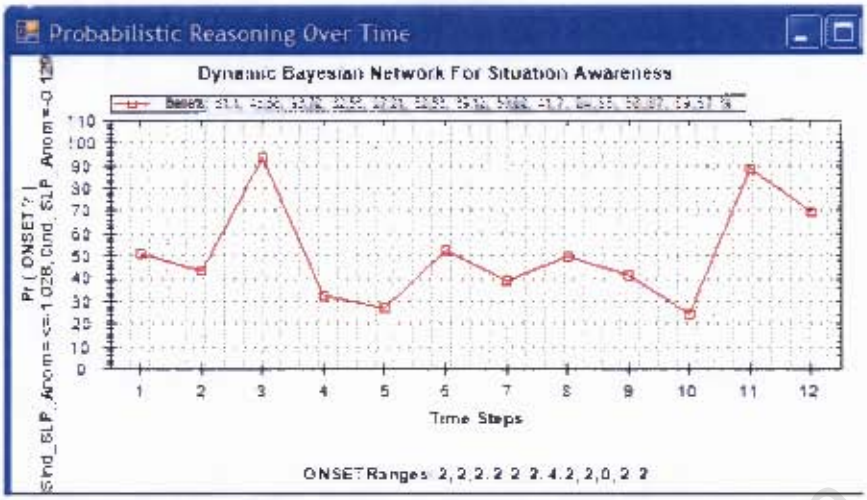


Figure 20a: Emergent Situation of *Onset* when *Sind_slp_Annom* <= '**-1.082**', and *Cind_slp_Annom* falls within '**-0.129<=2.47**'

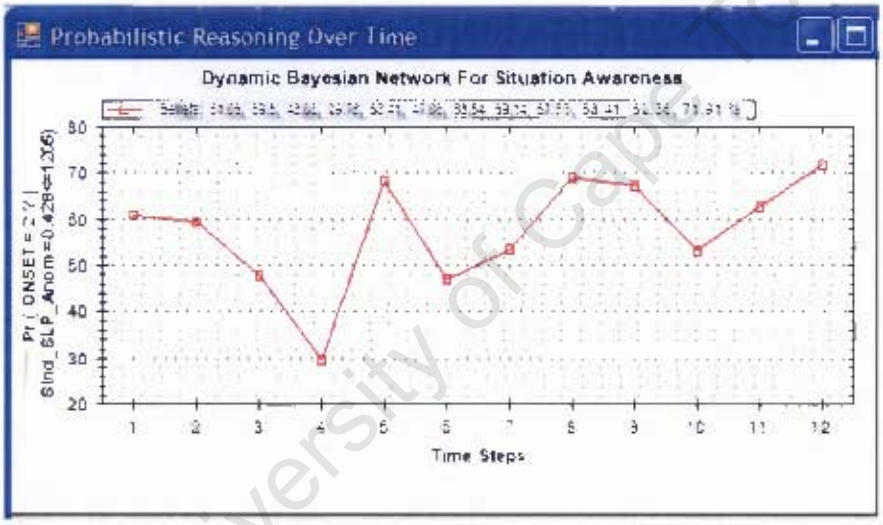


Figure 20b: Emergent Situation of normal onset when SInd SLP Anom is High (**SInd_SLP_Anom: 0.428<=1.205**).

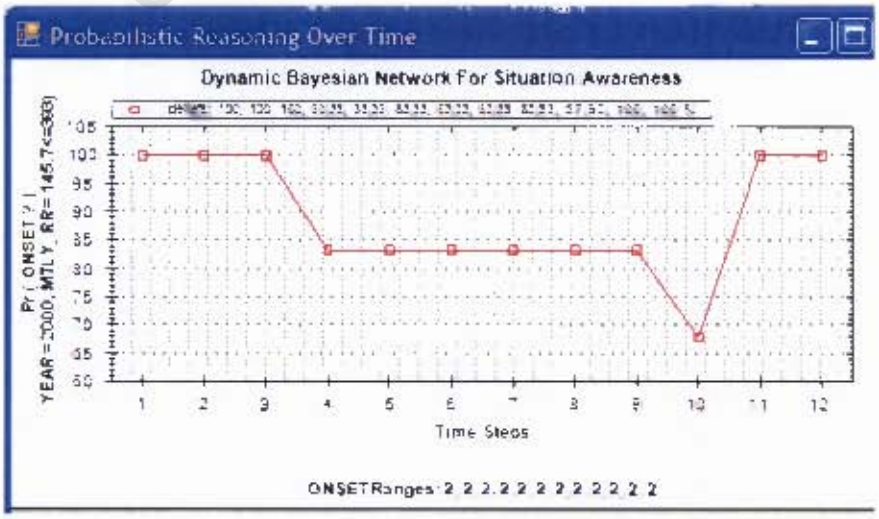


Figure 20c: Current Situation of **onset** for year **2000** when it is **very wet** (145.7<=333).

Q1: What is happening?

A1: From the graph (figure 20a), negative anomalies of the sea level pressure in South Indian Ocean, Macarene area, combined with near normal SLPs in Central Indian Ocean lead to high probability of normal onset of rainfall in Botswana in November (90%), while it is 70% in December for normal onset.

However, when we have positive SLP anomalies (**SInd_SLP_Anom: 0.428<=1.205**) in Mascarene area, degrees of beliefs for normal onset reduce for November (62%) and increase for December (72%).

In a wet year, like 2000 (figure 20c), degrees of beliefs for normal onset in November and December are 100%.

Q2: Why is it happening?

A2: Changes in sea level pressure in Southern Indian Ocean (Mascarene area) and Central Indian Ocean, have effect on degrees of beliefs for onset of rainfall and consequently on rainfall in the interior of Southern African continent. In a *wet* year, rains will most certainly set in the month of November (fig. 20c). The situation is similar for a *normal* year like 1998, (see fig. 23)

Q3: What will happen next?

A3: It is evident that as long as atmospheric changes such as the SLP anomalies set in, the rainfall onsets and their degrees of beliefs will be affected, as can be seen from figures 20a and 20b. If a *normal* or a *wet* year is forecasted, there is high degree of belief that rains will set in November.

Q4: What can I do about it?

A4: Weather forecasters, agricultural researchers and other stakeholders should look out for signals such as SLP anomalies in Indian Ocean that may affect the onset of rainfall. In most cases, there are higher degrees of belief for normal onset of rains in November, especially during *normal* and *wet* years.

The above sample result of probabilistic reasoning over time step (months) shows the degrees of belief for onset of rains given certain parameters, in this case

values of SLPs in Indian Ocean. However, as indicated in footnote 4, value of Onset parameter for a given year was entered from July of that year to June the following year. In this work we shall be interested mainly on the months when the rainy season starts in Southern Africa, (October - December).

4.7 Conclusion

In this chapter we described data sources and explained methodology used to analyse rainfall data to come up with Start of Season (SoS)/onset dates of summer rainfall.

We have also described the use of Principal Component Analysis (PCA) to delineate Botswana into three regions with similar rainfall characteristics in October, November and December (OND). Linkages between rainfall in the delineated homogeneous rainfall zones were then investigated through correlation analysis. Regions in global oceans were identified that had significant correlations.

Sea surface temperatures and sea level pressures were extracted from the oceanic regions, described above, that neighbor Southern African. These and other parameters were organized for use for analysis with Emergent Situation Awareness for dynamic Bayesian networks.

Chapter 5: Results and Discussions

5.0 Introduction

In this chapter we summarize results from statistical analysis and DBN for Emergent Situation Awareness (ESA) technology. The resulting DBN that represent temporal dependencies among weather parameters/climatic indices and onsets of summer rainfall in Botswana is presented in this chapter.

Using outputs from ESA, we also discuss the relationships of onsets of rainfall to various weather parameters and indices and mapping these parameters to synoptic features.

5.1 Results from Statistical Analysis - Rainfall onset

Results from the analysis show that the mean onset date of rainfall in Botswana is 25th November (day 148).

The earliest onset date is 1st October around the hilly areas around Kanye and Molepolole, and the latest is 23rd March in dry areas of Kgalagadi District, around Werda. Below is a plot of average onset dates (number of days from 1st July) for Botswana.

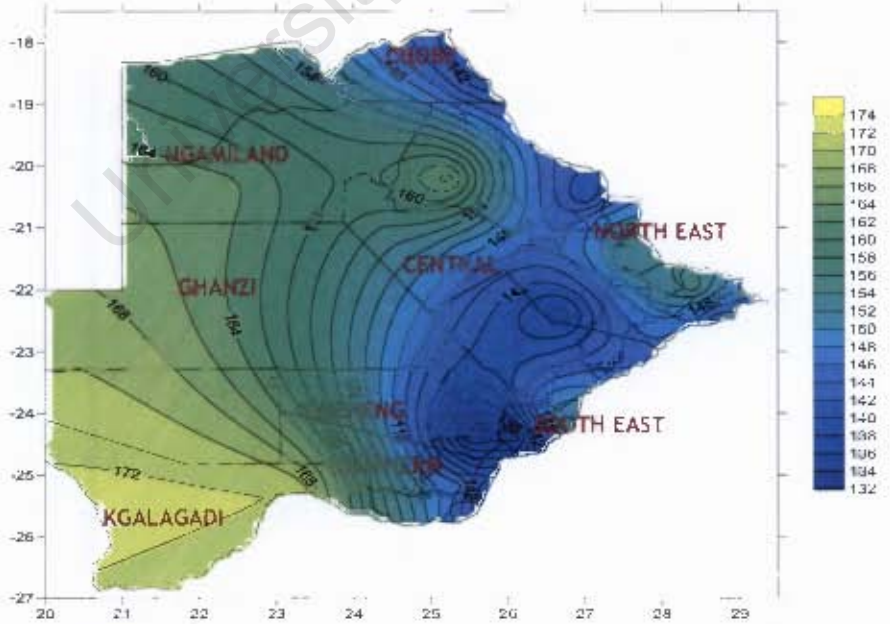


Figure 21: Spatial distribution of onset of rainfall in Botswana. Figures represent days after July 1st.

The earliest isolated onset is on South Eastern hilly areas around Otse on average 9th November. Lobatse station also recorded an earliest onset of day 93 (1st October 1993). Since Otse and Lobatse are on hilly areas, an early onset would suggest an orographic effect on arrival of moist easterly winds.

As expected, the southward march of ITCZ triggers onset of rains in the northern areas around Kasane and Pandamatenga on November 16th , spreading to Central, North-Eastern and South Eastern areas around Good Hope by 20th November. Marginal areas around Bobonong and Olifants Drift also receive rains from the middle of November, while rains set in through much of Central Kalahari by middle of December. The arid extreme south-west (Kgalagadi District) represented by stations Ncojane and Werda are the last to receive rains on 17th and 19th December respectively.

5.1.2 Variability of onset of rains

There is a high variability of onset of rainfall in Botswana. The average variability (standard deviation) of start of rains in Botswana is 38 days. The largest variability is found in the arid Kgalagadi District: Ncojane has variability of 55 days while Werda has 52 days. These are followed by semi-arid areas of Central District, mainly Orapa (49 days) and Rakops (43 days). The lowest variability occurs around Tutume in North-East and Pelotshela in the Southern District (20 days). Tadross (2005) noted that onset variability is partly forced by synoptic conditions over the region.

5.2 Results from Bayesian Network Analysis

Figure 22 below shows evolved dynamic Bayesian networks. These Directed Acyclic Graphs model temporal dependencies among weather parameters and climate indices. The BN nodes are:

a. *Station* - Rainfall Station
b. *Year* - Year (1971 - 2002)
c. *Mthly_RR* - Monthly Rainfall (in mm)

d. Ind_SST_Anom -Indian Ocean SST Anomalies[5] (^{0}C)
e. Atl_SST_Anom - Atlantic Ocean SSTAnomalies (^{0}C)
f. SOI - Southern Oscillation Index (Indicator for El-Niño/La-Niña) - (Warming/cooling episodes in Eastern Pacific)
g. (T- 3)SOI - three months lag in SOI
h. Atl_SLP_Anom - Atlantic Ocean Sea Level Pressure Anomalies (hPa/mb)
i. SInd_SLP_Anom - Southern Indian Ocean (Mascarene area) Sea Level Pressure Anomalies (HPa/mb)
j. CInd_SLP_Anom - Centra Indian Ocean (North of Malagasy area) Sea Level Pressure Anomalies (HPa/mb)
k. 500HPa_Anoms - 500 hPa hectopascal level anomalies (in meters)
l. **700HPa_U** - 700 hPa level *zonal* winds (meters per second)
m. **700HPa_V** - 700 hPa level *meridional* winds (meters per second).
n. **Onset** - Onset of rains in Botswana

The learning used a genetic algorithm in time-steps of *months*. The result was 12 frames (Frames 1 - 12) shown on figure 39 in Appendix 3. Frame 1 is a BN for January, while Frame 12 corresponds to December. Each Frame reveals the hidden relationships between the nodes mined from the data for that particular time step (month). The varying relationships are expected due to changes in rainfall over time.

Figure 22 below, shows selected Frames 1, 2, 11 and 12 from the Dynamic Bayesian network evolved for Botswana rainfall and associated weather parameters and climate indices. This is resulting DBN that represent temporal dependencies among variables (year, weather parameters/climatic indices and onsets of summer rainfall).

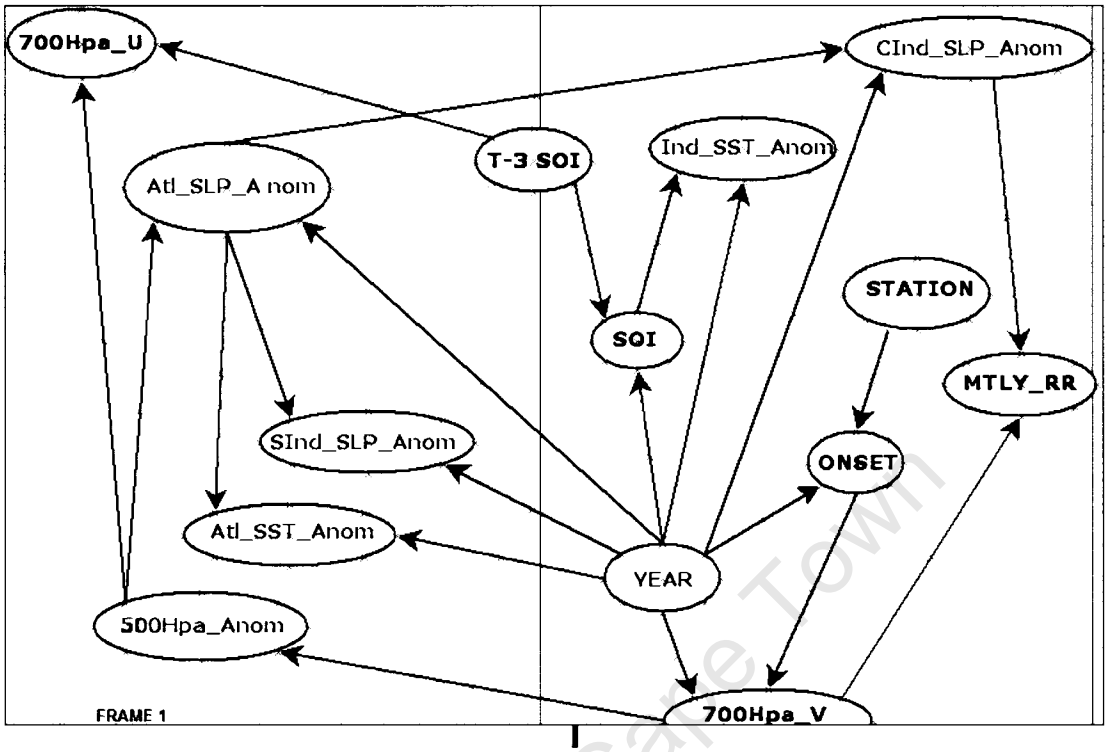
All through the four frames, it is clear that the ***Onset*** parameter is *characteristic* to any geographical location and year, that is, ***STATION*** **and** ***YEAR*** parameters have direct influence on ***Onset*** of rainfall. In Frame 1, the ***Onset*** of rainfall has direct

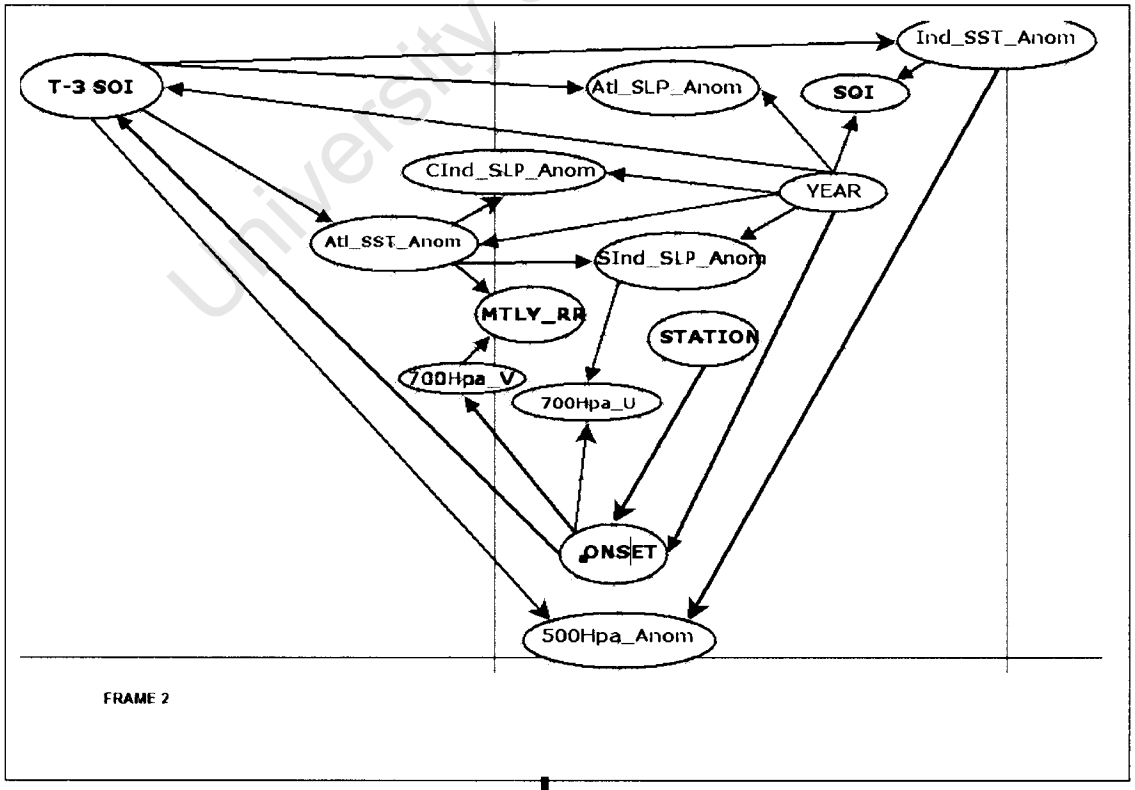
[5] Anomalies - Departures from the normal values

influence on meridional component of 700 hPa winds (***700 hPa_V***), which has direct influence on monthly rainfall (***MTLY_RR)***.

However Frame 2 shows that the ***Onset*** of rainfall has 'direct influence' on (***T-3)SOI***; zonal and meridional components of 700 hPa winds ((***700 hPa_U***) and (***700 hPa_V***). From the third frame, Frame 11, only ***SOI*** has direct influence on ***Onset***, while ***Onset*** parameter has direct influence on monthly rainfall (***MTLY_RR)***.

In the last frame, Frame 12, the ***Onset*** parameter has direct influence on ***(T-3)SOI***, the three months lead time of Southern Oscillation Index (SOI). Using "forward and backward" propagation concept, the interpretation is that ***(T-3)SOI*** has direct influence on the ***Onset*** of rainfall in Botswana. Both zonal and meridional components of 700 hPa winds ((***700 hPa_U***) and (***700 hPa_V***) have direct influence on rainfall monthly rainfall (***MTLY_RR)***, similar to Frame 2.
Figure 39 in Appendix 3 contains the complete set of Frames 1 - 12 of the evolved DBN representing all the months.





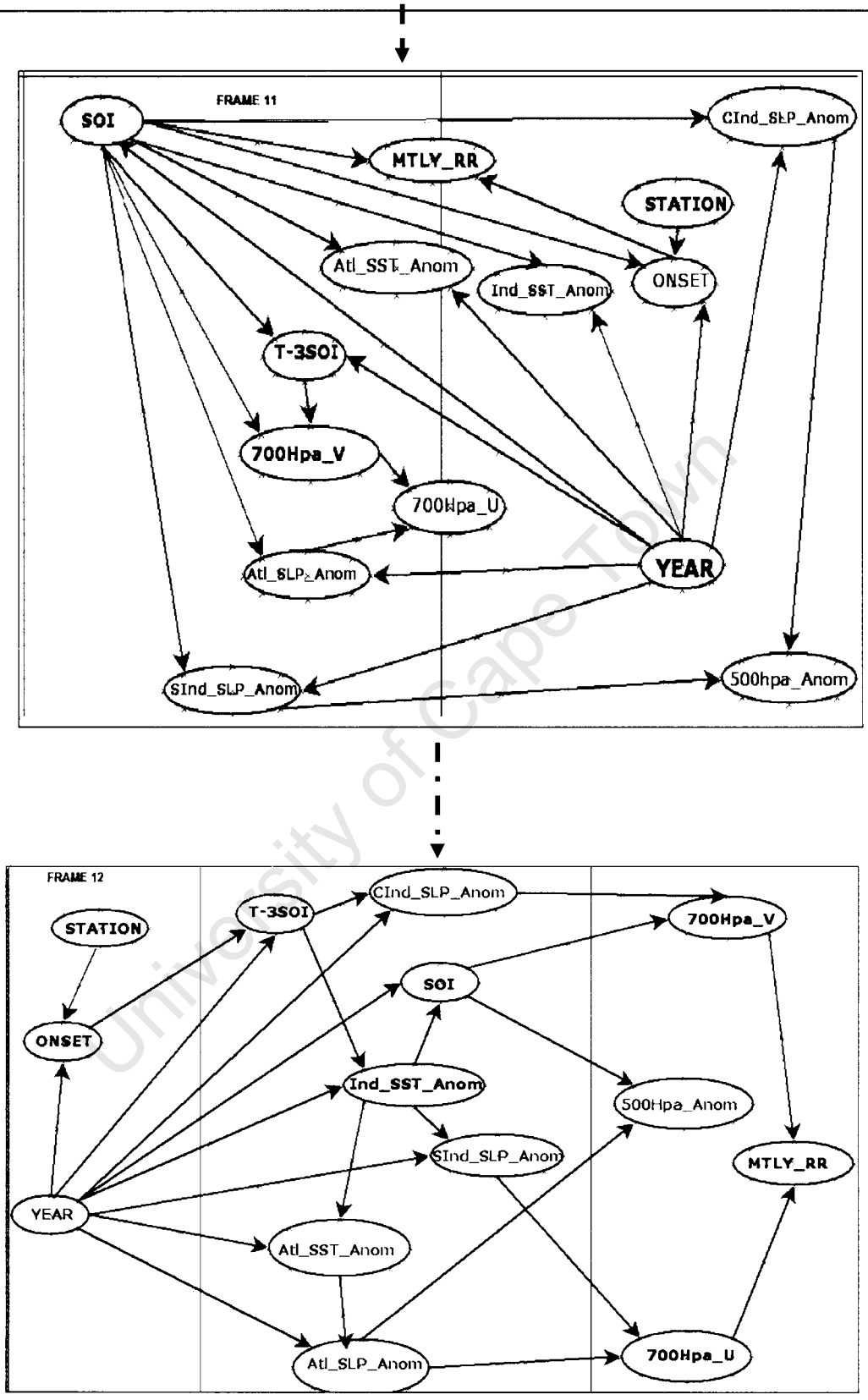


Figure 22: Dynamic Bayesian network of Botswana rainfall and associated weather parameters and Climate indices

The direction of the links between any two parameters or their positions apart, directly or indirectly, does not affect reasoning of the model. It is in agreement with the "forward and backward propagation" concept.

Most existing DBN techniques assume that frames do not change over time. Most of the techniques described before are based on the simple DBN, a Hidden Markov Model, and have been used for tasks of describing time varying patterns. These techniques do not capture all variability that we are assessing (Murphy & Mian, 1999). The ESA technology is is suitable for tasks that change over time, and can furthermore be applied to complex and ad-hoc algorithms whose designs are determined by the knowledge of expert users (Mihajlovic & Petkovic, 2001).

5.2.1 Relationships of rainfall onsets to various parameters

5.2.1.1 Normal and Wet years

We look at the year 1998, a normal rainfall year in Botswana that also had a normal onset of rains from statistical analysis. The national rainfall index was 102 and nearly all rainfall stations showed normal onset. Figure 23 below shows a graph of probabilistic reasoning over time followed by a User's interaction. We will omit the user's interactions in subsequent graphs.

The graph shows probabilities of normal onset of rains in year 1998 for the month time steps (1 to 12 representing January to December) given that monthly rainfall (MTLY_RR) was at least 56.2mm but more than 26.6mm. This year was chose since it was a normal rainy year in Botswana. We can see that probability for a normal onset for November (when the season starts in Southern Africa) is high (90%), subsequent months are equally high, November to March reveal very high probability of onset of rainfall (above 82%). However we have lower degrees of belief for normal onset for April to October (less than 70%).

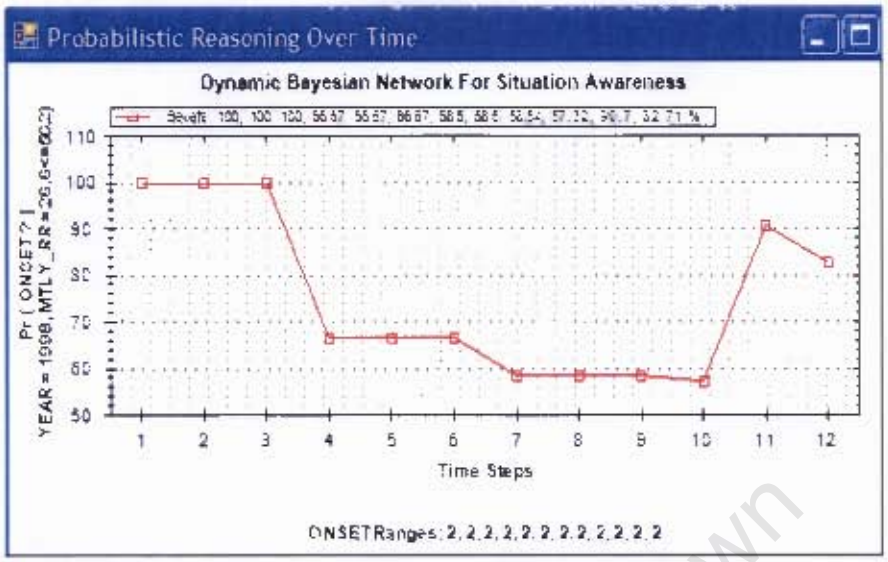


Figure 23: Emergent Situation of Onset in **1998** when monthly rainfall (*MTLY_RR*) was Within '**26.6 <= 56.2**'. The Onset was normal (2).

Here is a User's Interaction with ESA

Q1: **What is happening?**

A1: From figure 23, the onset of rainfall in *1998* is generally *normal* within the monthly rainfall range but with varying degrees of beliefs. **November** to **March** reveal very high probability of onset of rainfall (above 82%), given that monthly rainfall fell between 26.6 and 56.2 mm. **April** to **June** is reasonably normal (about 70%) while **July** to **October** is low normal onset (about 60%).

The high probability of normal rainfall onset in the month of November (90%) confirms the statistical result that on average the onset of the rains in Botswana is in the month of November.

Q2: **Why is it happening?**

A2: It has been observed in Botswana that the rainy season falls between **November** and **March**. There is little rainfall from **May** to **September** but, according to DBN for ESA, the actual dates and the onset degrees are unknown, as evident from figure 23.

Q3: **What will happen next?**

A3: If there is *significant* change in climate, most of these results from figure 23 will obviously change.

Q4: What can I do about it?

A4: Using results from figure 23, Water conservation and agricultural activities should take these onset degrees into consideration.

A similar scenario is witnessed for individual *very* wet years: 1974 and 2000. For the year 1974 probabilities of normal onsets are: October (60%), November (100%) and December (100%), while for year 2000 they are October (65%), November (100%) and December (100%).

An exception was 1977 which was a wet year, but experienced late onset. It has a reduced degree of belief. Probabilities for normal onset are: October (40%), November (60%) and December (80%). In this case of 1977, the onset could have taken place after December.

The year 1985 was a dry year, and had a late onset in almost all the areas, except Serowe in Central District that had failed onset. From figure 25 below, the probabilities of failed onset given that the monthly rain (MTLY_RR <=26.6) and onset was *failed* (4) are: October (30%), November (65%) and December (55%) respectively.

Another dry year was 1993. Most stations recorded a *normal* onset, though nationally the onset was early (1). The probabilities of *normal* onset given that the monthly rain (MTLY_RR <=26.6) and onset was *normal* (2) are: October (70%), November (60%) and December (85%) respectively.

From the resulting DBN, figure 22, the *monthly rains* (MTLY_RR) are ***caused*** by the following variables (weather parameters and indices): Onset of rains (*Onset*); the Southern Oscillation Index *(SOI); zonal and meridional winds at 700hPa level (700hPa_U and 700hPa_V); Atlantic Ocean Sea Surface Temperature anomalies (Atl_SST_Anom) and Central Indian Ocean Sea Level Pressure Anomalies (CInd_SLP_Anom).*

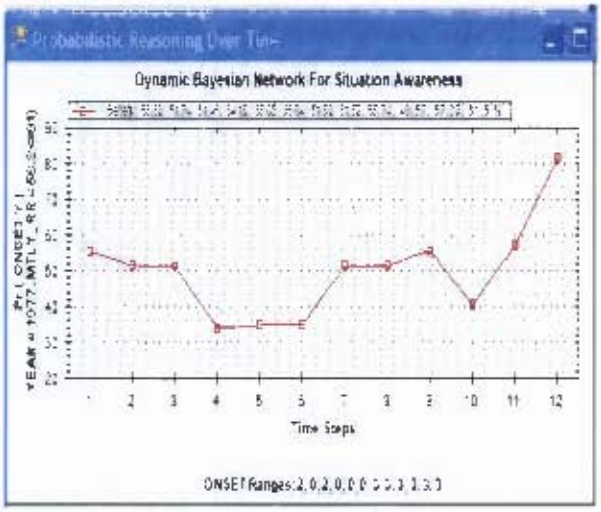


Figure 24: Emergent Situation of *Onset* in 1977 when it is wet (56.2< 91)

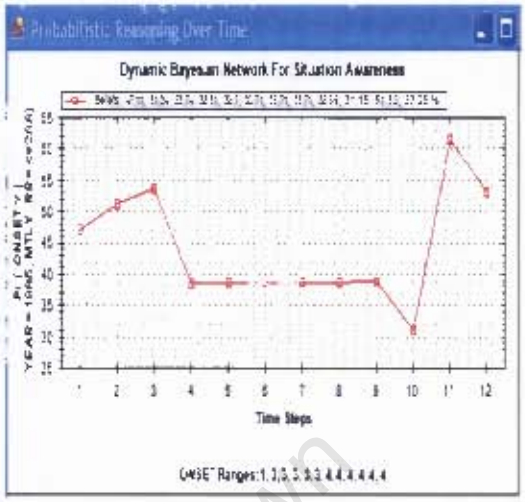


Figure 25: Emergent Situation of *Onset* in 1985 when it is Dry (<=26.6)

In a dry year, some of the large scale weather systems that contribute to sustained rainfall activities will either be not fully developed (Mascarene High and moist easterly winds); misplaced (position of ITCZ too far to the north or very diffuse) or shifted (St Helena High and associated ridging elongated towards the sub-continent) (Washington & Todd, 1999).

5.2.1.2 El-Niño/Southern Oscillation Index and Onset variability

The figures 26 to 28 below give graphical result for DBN for Situation Awareness for late and false onsets given when the SOI is '-3.8<=0.9' (near normal conditions in the East Pacific or mild *El-Niño* condition) and T-3SOI is high (6.9 <=18.5), (*La Nina* episode). We see that for mild El-Niño, the probabilities are: October (35%) for early onset; November (60%) for normal onset; and December (85%) for normal onset.

This result is reflected in Frame 11 (temporal BN for November), figure 22, where we saw that the Southern Oscillation Index ***(SOI)*** variable has direct influence the ***Onset*** and other variables (***MTLY_RR, Atl_SST_Anom, Atl_SLP_Anom, Ind_SST_Anom, Cind_SLP_Anom, SInd_SLP_Anom***, ((***T-3)SOI, 700 hPa_V***).

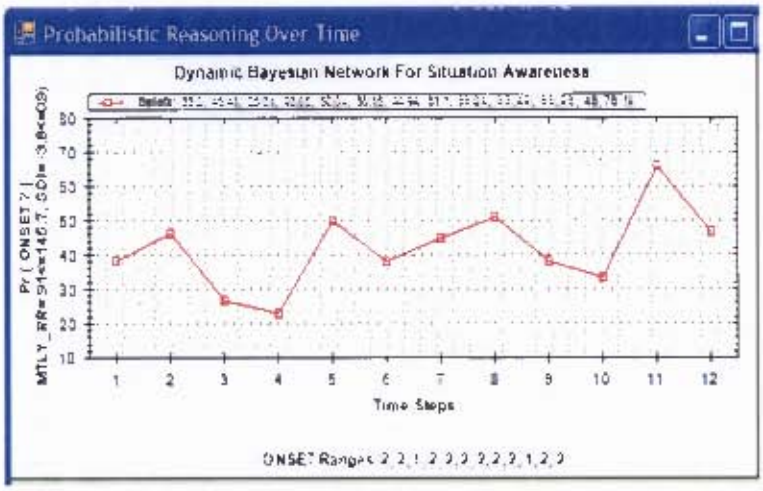


Figure 26: Emergent Situation at *Onset* when *Mthly_RR* falls within '91<=145.7' and *SOI* falls within '-3.8<=0.9'. Case for weak El-Nino and near-normal conditions

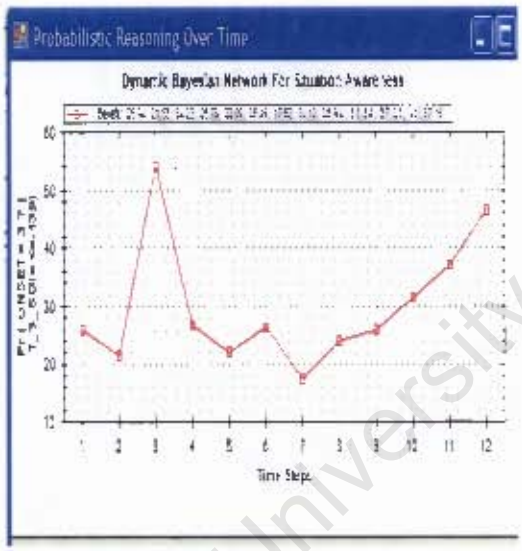


Figure 27: Emergent Situation of late onset when *(T-3) SOI* is small (<= -13.9). (Well marked El-Nino situation).

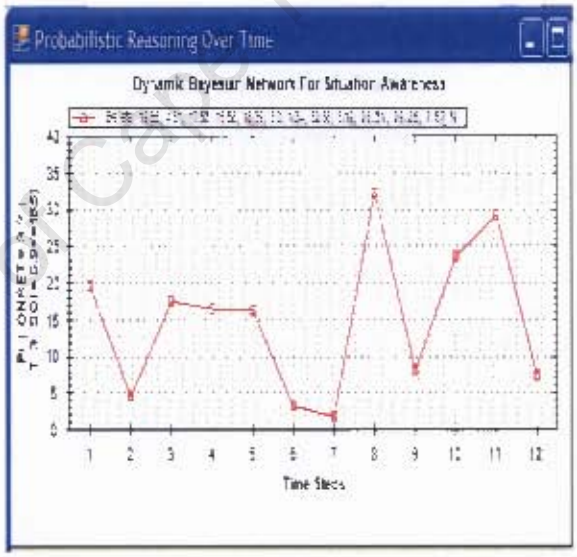


Figure 28: Emergent Situation of late onset when *(T-3) SOI* is high (6.9<= 18.5). (Well marked La-Niña situation)

When we have a well marked El-Niño episode, as in figure 27 above, ((T-3)SOI <= -13.9), there is some marked probabilities for a late (3) onset in November (40%), December (47%) and March (55%).

Whenever there is a fully established *La-Niña condition* there is a very low probability (less than 30%) to have late and false onsets. An example is the figure 28 above, where parameter (T-3) SOI is high (6.9<= 18.5). Whenever such an episode occurs with three months lead time, corresponding atmospheric features

would be set up (Nicholson et al., 2001). These features are conducive for reliable and sustained rainfall following the onset.

5.2.1.3 Sea Level Pressure and Sea Surface Temperature anomalies

Sea Level Pressure

From figure 20 and accompanying discussions, the probability for normal onset given the negative SLP anomalies in the South Indian Ocean (Mascarene region), (SInd_SLP_Anom <= -1.082) are: October (30%), November (90%) and December (50%). However given positive pressure anomalies (SInd_SLP_Anom 0.4<= 1.25), the probabilities for normal onset are higher: October (50%), November (62%) and December (70%). The strengthening of Mascarene High pressure cell to the south of Madagascar induces low level moisture flux along the east coast of Southern Africa that triggers wet spells into the interior of the sub-continent, (Todd et al. 2004). This possibility leads to a higher degree of belief for normal onset of rains.

Sea Surface Temperature

The result below for warming and cooling of sea surface in Indian Ocean shows that probabilities for normal onset are higher (lower) given anomalous warming (cooling) of SSTs. This agrees with arguments advanced by Reason and Mulenga (1999). The pattern is similar for Atlantic Ocean.

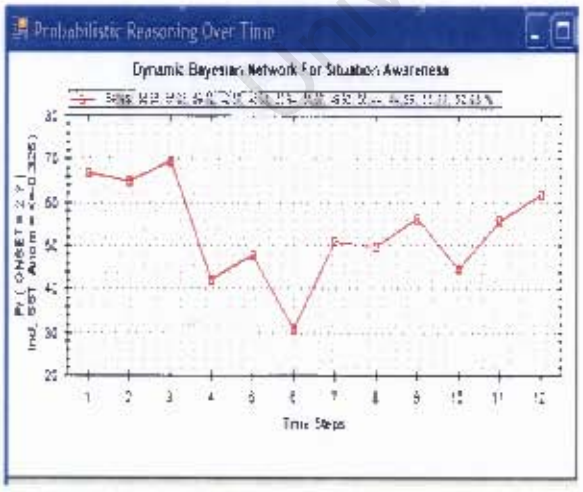


Figure 29: Emergent Situation of **normal onset** when *Ind_SST_Anom* is **small** (<= - 0.325).

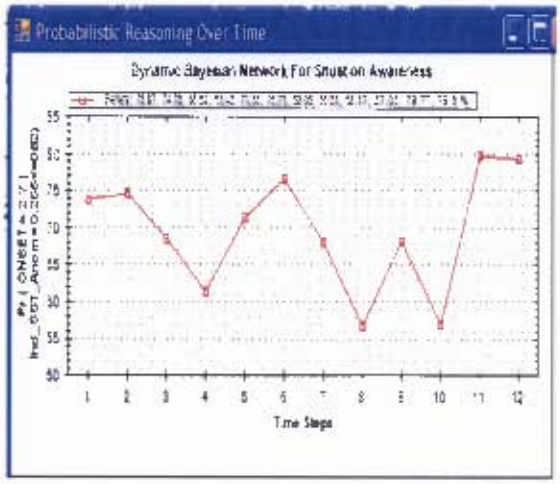


Figure 30: Emergent Situation of **normal onset** when *Ind_SST_Anom* is **high** (0.265<=0.82).

5.2.1.4: The 500 hPa geopotential height anomalies

One of the parameters that best relate to the onset of rainfall is the 500hPa geopotential anomalies (500HPa_Anoms). Positive anomalies over the interior of the continent, (around Botswana), points to a high pressure system in medium levels (700 hPa to 500 hPa). This may signify dominance of "Botswana Upper High" pressure system. This system is associated with subsidence, downward movement of air that suppresses convection and development of clouds (Preston-Whyte, 1999). Dry conditions are associated with this system. Isolated rains would only take place around hilly areas where there is possibility of orographic effect. See figure 36 for NCEP/NCAR re-analysis of 500 hPa geopotential height anomalies and 700hPa meridional wind anomalies for a sample dry year (1985) and wet year (2000).

Conversely, negative geopotential height anomalies may point to a deep low pressure system extending from the surface to medium levels, which is often associated with ITCZ, westerly trough and its associated extra-tropical depression[6] extending from the Cape into the sub-continent. These features, earlier described, give rise to enhanced tropical convective activities, given sufficient moisture and favorable stability conditions. Note from the figure 37a-b below that there were negative 500 hPa geopotential anomalies of between 5 to 10 meters in the year 2000 (Wet year), and near normal or zero in 1985, (dry year). This is in agreement with (Matarira, 2006) who observed that stronger pressures over the sub-continent compared to oceanic areas, leading to positive pressure gradient force, is characteristic of dry events, and vice-versa.

The result, as shown in the figures 31 and 32, from the ESA for normal onset when there are positive anomalies (4 <= 65 m) reveal that there is reduced degree of beliefs, compared to that of negative anomalies (500HPa_Anoms <= -7.974). The probabilities of onset for the negative anomalies of geopotential heights are: October - 40%; November - 80% and December - 55%. As before, this agrees with results of other parameters to a normal onset in the month of November.

[6] Extra-tropical depressions are intense low pressure systems often accompanied by cold or warm frontal systems.

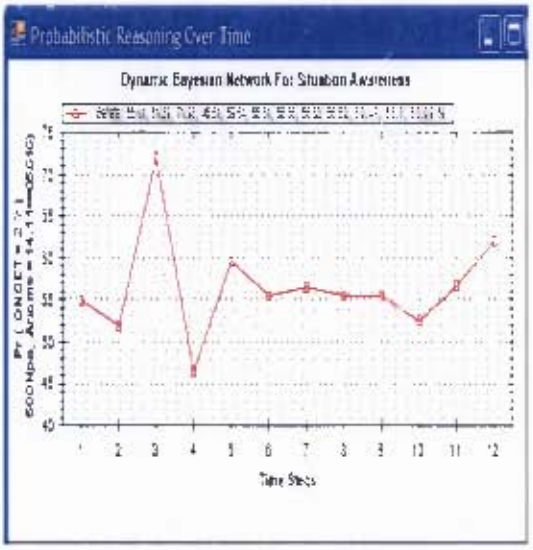


Figure 31: Emergent Situation of **normal onset**, when **500HPa_Anoms** falls within **4.11<=65.616**

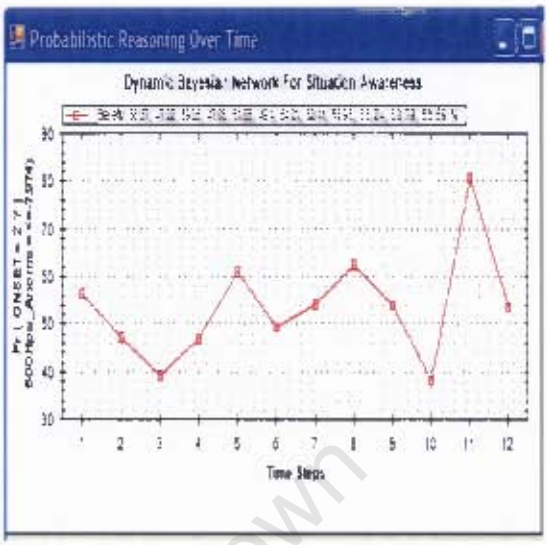


Figure 32: Emergent Situation of **normal onset**, when **500HPa_Anoms <= -7.974**

Using a similar parameter, the 500 hPa geopotential height anomalies, Tadross (2005) found that early onset of rainfall further north in Zimbabwe was accompanied by increased frequency of 500-hPa positive geopotential height anomalies to the South-East of the continent. Depending on the position of the anomalies, they may affect the lower atmosphere by strengthening the Mascarene High which then advects moisture from SWIO over the continent.

5.2.1.5 Wind systems at 700 hPa level

Directions and wind speeds of winds at 700 hPa are indicative of the circulation regimes that support various rainfall patterns; these are the ITCZ (Okoola, 1999), and extra-tropical troughs and associated interactions with tropical systems. From the ESA results, for westerly winds (700HPa_U <- -3.244) there is a higher probability of a normal onset of rains for September (65%) and lower for November (50%), while for December, the probability is 60%. However the probabilities for normal onset for strong easterly winds (700hPa_U: 3.337<=8.938) of between 3 to9 meters per second that advect moisture from Indian Ocean are: October (46%), November (60%) and December (60%).

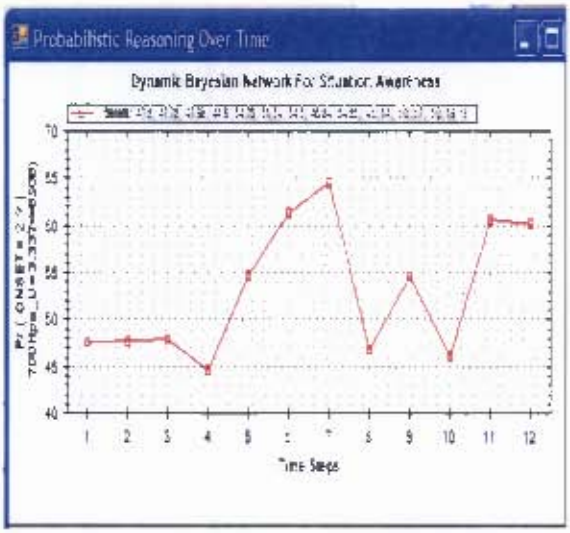


Figure 33: Emergent Situation of **normal onset**, when **700HPa_U** falls within **3.337<=8.938**

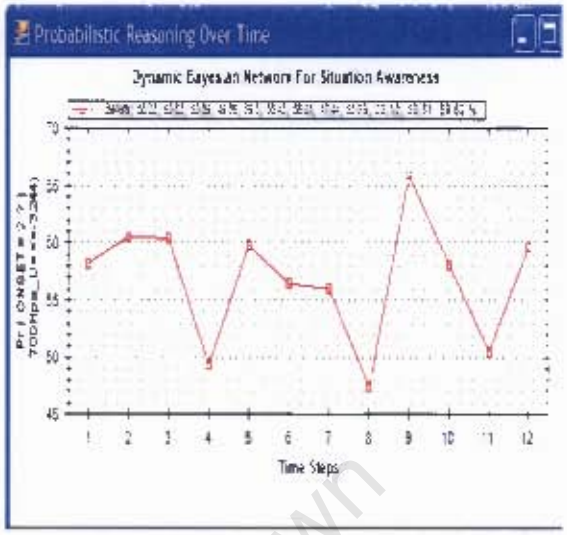


Figure 34. Emergent Situation of **normal onset**, when **700HPa_U <= -3.244**

Figure 35 below suggest weak southerly winds, (700HPa V<= -1.904), over the region that may be indicative that the ITCZ may be located further north of Botswana. As in the case of geopotential height anomalies at 500 hPa, southerly flow at 700 hPa may also be associated with extra-tropical troughs and interactions with tropical systems. In this case there are higher probabilities for normal onset: October (47%), November (65%) and December (75%). From figure 37c-d, there was strong northerly component of 700 hPa windflow in 1985 (a dry year), compared to strong southerly flow in year 2000 (a wet year).

On the other hand when we have northerly flow of about 1 to 5 m/s, figure 35, (700hPa_V: 0.885<=4.58), probabilities for normal onset are reduced, they are October (50%), November (60%) and December (55%).

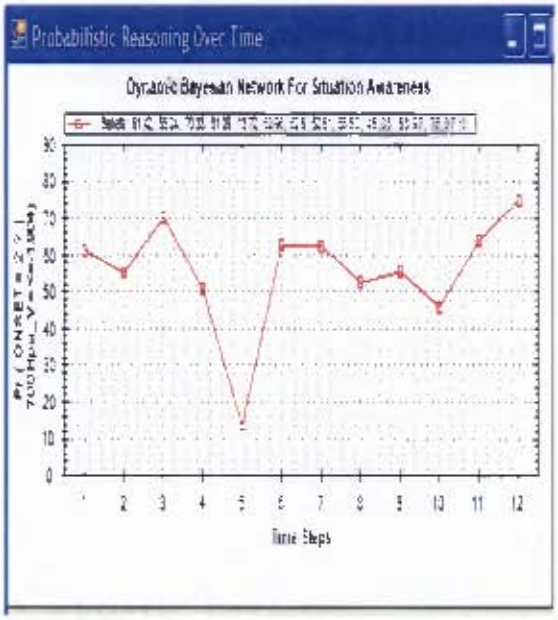


Figure 35: Emergent Situation of **normal onset**, when **700HPa_V<= -1.904**

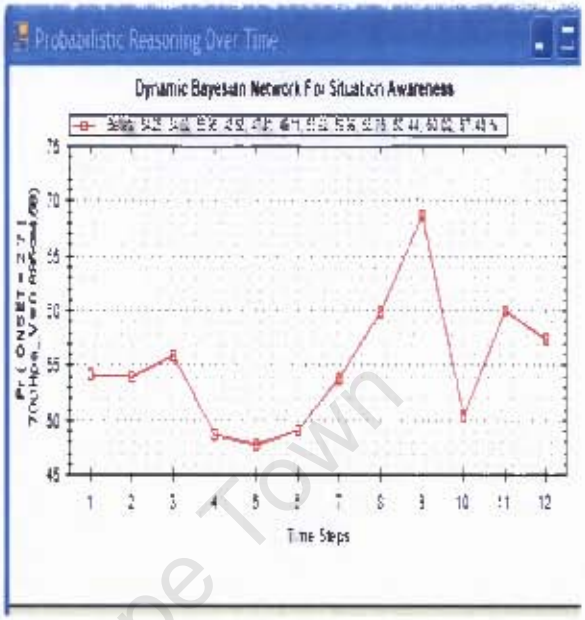
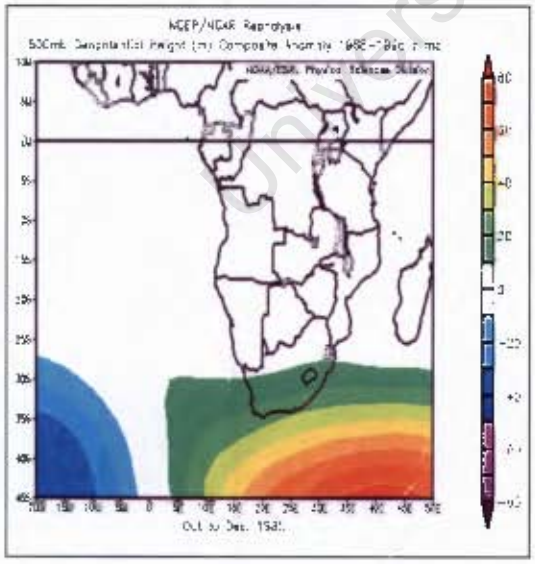
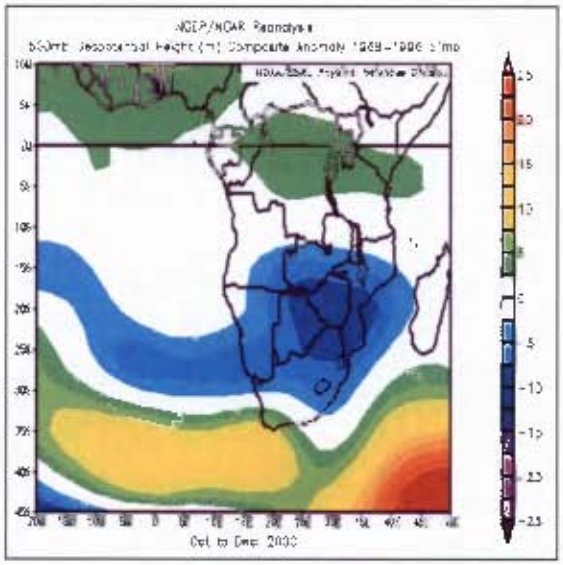


Figure 36: Emergent Situation of **normal onset**, when **700HPa_V** falls within **0.885<=4.58**

NCEP/NCAR reanalysis for 500 hPa geopotential height anomalies

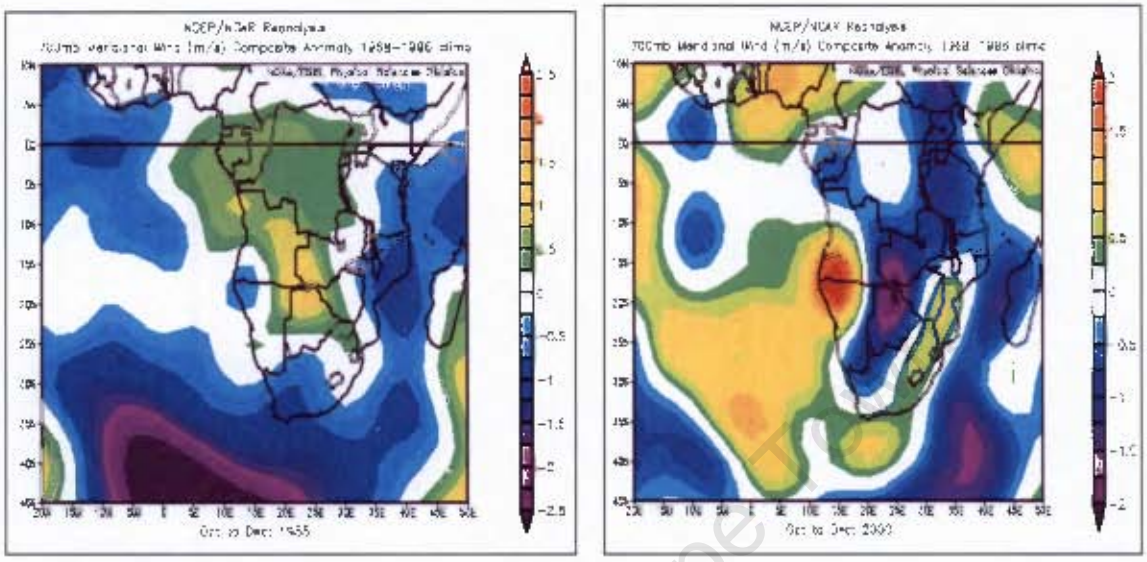


(a) Case for Dry year - 1985



(b) case for wet year - 2000

NCEP/NCAR reanalysis for 700 hPa meridional wind anomalies



(a) Case for Dry year - 1985 (b) case for wet year - 2000

Figure 37: NCEP/NCAR reanalysis for (a), (b) 500 hPa anomalies and (c), (d) 700 hPa meridional wind anomalies for 1985 (dry) and 2000 (wet) years respectively.

5.3 Summary and conclusions

In this chapter, results for statistical and DBN for Emergent Situation Awareness are presented. The resulting DBN that models temporal dependencies among weather parameters/climatic indices and onsets of summer rainfall in Botswana has been presented. The model evolves over time and reveals what is currently happening in a domain of interest, in this case relationships between onset of rainfall and various parameters and degrees of beliefs of those onsets given varying measures of parameters.

Using outputs from ESA, we also discussed the relationships of onsets of rainfall to various weather parameters and indices and mapping these parameters to synoptic features. We also discuss causal relationships in the DBN.

Results from DBN for ESA showed that in the years that had normal and above normal rainfall (wet), the probability of a normal onset was high. The month of November had probability of 90% for a normal onset. This agreed with statistical

results that showed that the mean onset date in Botswana is 25th November. The mean variability (standard deviation) of onset of rains in Botswana is 38 days.

a. During an El-Nino year, there is marked probability for late onset of rains: 40% in November and 55% in March. During a La-Nina year there is low probability for late and false onset of rains. This agrees with the findings of Nicholson & Selato (2000) and Semazzi & Indeje (1999) that La-Nina episode in Southern Africa is associated with anomalously above normal rainfall. This scenario then would resemble (a) above.
b. Whenever we have negative sea level pressure (SLPs) anomalies in South East of Southern Africa (Mascarene area) and near normal SLPs in Central Indian Ocean, degrees of beliefs are still high for normal onset of onsets for the months of November (88%) and December (70%). However Todd et al. (2004) noted that the strengthening of Mascarene High pressure cell induces moisture influx into the interior of Southern Africa, leading to wetter conditions. It was expected then from the ESA that negative anomalies would lead to marked reduction in probabilities for normal onset.
c. There are higher degrees of belief for normal onset of rains given positive (warming) sea surface temperatures (SSTs) anomalies in Indian Ocean, while there are reduced degrees of belief for negative (cooler) SSTs. This agrees with reasons advanced by Mulenga (1999).
d. Positive 500 hPa geopotential height anomalies in the interior of the sub-continent, signifying Botswana Upper High feature, lead to lower degrees of belief for normal onset of rains. On the other hand negative 500 hPa geopotential height anomalies leads to higher degrees of belief for normal onset for the month of November (80%), but lower for other months (less than 50%). Tadross (2005) found that early onset of rainfall in Zimbabwe was associated by frequent 500 hPa geopotential height anomalies in the South East of the continent.
e. A marked southerly component of wind flow at 700 hPa level lead to higher probabilities for normal onset, while a northerly flow lead to reduced probabilities for normal onset. The results showed that during wet years, for example the year 2000, there was marked southerly flow across Botswana. Conversely in a dry year, 1985, we had a marked northerly flow over the region.

f. From the DAG that models relationships among the parameters, the monthly rains (MTLY_RR) are caused by the following parameters: Onset of rains (Onset); the Southern Oscillation Index (SOI); winds at 700hPa level (700hPa_U, 700hPa_V); Atlantic Ocean Sea Surface Temperature anomalies (Atl_SST_Anom) and Central Indian Ocean Sea Level Pressure Anomalies (CInd_SLP_Anom).

g. The Southern Oscillation Index (SOI) variable has direct influence the Onset and other variables (MTLY_RR, Atl_SST_Anom, Atl_SLP_Anom, Ind_SST_Anom, Cind_SLP_Anom, SInd_SLP_Anom, ((T-3)SOI, 700 hPa_V) for four months (time slices), figure 22.

h. Similarly from figure 22, the three month lead time in SOI ((T-3)SOI) has direct influence, for the four time slices, on the following parameters:

i. DBN for January - SOI and 700hPa_U

j. DBN for February - 500hPa_Anom, Ind_SST_Anom, Atl_SLP_Anom and Atl_SST_anom.

k. DBN for November - 700hPa_V

l. DBN for December - Ind_SST_Anom, CInd_SLP_Anom

Chapter 6: Conclusion

The analysis of daily rainfall with statistical software yielded onset dates for each station and other statistics: mean, variance and standard deviation of onset dates. The ESA for DBN yielded the DBN that modeled temporal dependencies among weather parameters and graphs of probabilistic reasoning over time. As mentioned before, there are limitations and problems of modeling sequential data and multi-dimensional inputs with statistical and classical approaches to time series prediction. One may not incorporate prior knowledge with these classical approaches ((Murphy, 2002), Zukerman & Albretch, 2004). The DBN, a type of a state space model, describes a system that is dynamically (temporal) changing or evolving over time and enables users to monitor and predict future behavior of the system (Milhajlovic & Petkovic, 2001).

The Emergent Situation Awareness (ESA) for DBN reveals more information from complex models. It revealed what is currently happening over time (month time-step). Each of the parameters and climate indices revealed varying degrees of beliefs for early, normal, late or failed rainfall onsets in Botswana. The result from statistical analysis was however used for analysis with ESA. This work was limited to establishing degrees of belief of onset of rains and parameters that best indicate the onset of rains. The causal relationships in a DBN may not be feasible in statistics. There was no attempt to compare homogeneous zones obtained from PCA analysis with ESA for DBN.

Results from ESA show that the onset of rains takes place in the month of November. While ESA for DBN does not show the actual dates, since the time-step was *month* and not *daily*, the statistical methods captured actual dates of onset of rains. In normal rainfall years and wet years, the degrees of belief for normal onset are higher in November (as much as 90%). Similarly there was higher probabilities for normal onset in November (about 80%) whenever we had anomalous warming in Indian Ocean, *Ind_SST_Anom* (0.27 <= 0.82 ^{0}C); About 80% probability for negative anomalies of geopotential heights at 500hPa level (*500hPa_Anoms*<=*-7.97*m and between 65% to 75% probabilities for November and December for southerly component of winds at 700 hPa (*700 hPa_V*<= *-1.9 m/s*).

Statistical analysis of daily rainfall showed that average onset date for Botswana was found to be 25th November (day 148). The earliest being 9th November in hilly areas near Lobatse, spreading to central and northern regions by 20th November. The average latest onset is 19th December in the arid Kalahari in the south-west.

The average onset variability in Botswana is 17 days, the largest being in semi-arid Kalagadi in the west (55 days) and the least in South-East (38 days).

Wet years, like 1974 and 2000, reveal highest degrees of beliefs for normal onsets for November and December, (100%). It is expected that large scale weather patterns that sustain rainfall are in place: ITCZ, north-easterly and easterly monsoon winds, mid-latitude westerly trough, and sub-tropical high pressure cells (Mascarene and St Helena Highs) at appropriate positions, Washington & Todd (1999).

There are, however, mixed degrees of beliefs for dry years. Some dry years like 1985 have fairly large degrees of beliefs for *failed* onset for November (60%), while other dry years, like 1993, have large probabilities for *normal* onset for November (60%) and December (85%).

We have about 65% probability for normal onset when there are near normal conditions in the East Pacific (SOI: -3.8<=0.9) and monthly rainfall is high (MTLY_RR: 91<=145). Whenever there is a marked La-Niña episode, having high SOI values (T-3SOI: 6.9<= 18.5), we have a very low degree of belief (less than 30%) for late (3) and *false* (4) onset in the region. Most wet years in Southern Africa are associated with La-Niña phenomenon; there are higher degrees of beliefs for normal onsets. Marked La-Niña phenomenon is associated with above average rainfall in Southern Africa, (Nicholson & Selato, 2000), (Nicholson et al. 2001).

Conversely during an El-Niño episode (T-3SOI: <=13.9), there is higher degree of belief (above 30% in October and November, and 60% in December) for a *late* onset. As explained earlier, the anomalous warming of sea surface waters in East Pacific

and consequent shifting of convective zones from Indonesia to Central Pacific causes a reversal of east-west circulations and affects weather patterns globally.

This three-month lead time of Southern Oscillation Index (T-3SOI) is a good indicator that may be useful for prediction of onset of rainfall in Southern Africa. Weather Bureaus and Meteorological Services in the United States, East, and Southern Africa use ENSO indices to predict performance of rains in any season. See NOAA Climate Prediction Centers' quarterly bulletins and forecasts at http://www.cpc.noaa.gov/products/analysis_monitoring/enso_advisory. This climate index is so important that the international community now monitors weather, including SSTs and ENSO in the entire equatorial Pacific on regular basis through TAO/TRITON[7] project (available at website http://www.pmel.noaa.gov/tao).

There is a high (low) degree of belief for negative (positive) anomalies for normal onset for 500 hPa geopotential height anomalies over the region. The negative anomalies are associated with the marked surface low pressure system and convective activities and rains. Conversely, positive anomalies may mean a stronger Botswana Upper High pressure system, which is characterized generally dry weather.

Marked southerly winds at the 700hPa level over the sub-continent is an indicator of an extra-tropical trough and associated interactions with tropical systems in Southern Africa. In this scenario, we have high probabilities for normal onset of rains: October (47%), November (65%) and December (75%).

In this work, we have shown that ESA for DBN suggest that all the parameters and climate indices positively or negatively influence the onset of rains in Botswana. The differences are the degrees of beliefs. Owing to the complexity of the atmosphere and its processes, weather parameters and climate indices are related and interlinked. Several parameters determine the rainfall, as was seen from the resulting DBN and hence define the onset. However, the results in this work have

[7] The Tropical Atmospheric Ocean Project (TAO/TRITON) is a real-time data collection platform of moored buoys in Equatorial Pacific. This project is meant to improve detection, understanding and prediction of El-Nino and La-Nina.

shown that the following parameters revealed higher degrees of beliefs for onset of rains in Botswana:

- The three month lead time of Southern Oscillation climate index *(T-3SOI)*;
- Weather parameter, geopotential height anomalies at 500 hPa level *(500hPa_Anoms)*;
- Meridional component of wind flow at 700 hPa level *(700hPa_V)*.

Chapter 7: Future Work

Sea Surface Temperatures play a key role in driving atmospheric process in the globe, (Bony, Lau & Sud, 1997), (Lu & Dong, 2004). It is the oceanic component of El-Nino/Southern Oscillation (ENSO). In this project I mainly used the atmospheric component - the SOI and (T-3)SOI. I also used SSTs for Atlantic and Indian Oceans only. Future work to assess onset of rainfall in Southern Africa should use SSTs of all the Nino regions: Nino 1+2, Nino 3, Nino 3.4 and Nino 4 (see figure 4b).

This work used larger time-step, (the *Month).* Future work should use shorter time-steps, like 10 day *(dekad)* or 5 day *(pentad)* periods. The results could then be compared with statistical determination of *onset dates.*

Further exhaustive experiments should be carried out to determine degrees of beliefs for early, normal or late onset given monthly rainfall. This may answer the question that an early onset would predictably lead to a very good (wet) rainy season, and a late onset would necessarily signal below normal (poor rainfall) season in Southern Africa.

Chapter 8: References

1. Ati O.F.; Stigter C.J. & Oladipo E.O, (2002): A comparison of methods to determine the onset of the growing season in Northern Nigeria. *International Journal of climatology, 22, 731-742.*
2. Bazira, E., & Ogallo, L. A. (1999): "Potential of seasonal rainfall prediction over East Africa as derived from Sea Surface Temperature modes of neighboring Oceans". *J of African Meteorological Soc., Vol 4 No.1. Nov 1999.*
3. Basalirwa, C.P.K., Odiyo, J.O., Mngodo R.J., & Mpeta E.J. (1999). The climatological regions of Tanzania based on the rainfall characteristics. *International Journal of climatology*, Volume 19, Issue 1 , Pages 69 - 80
4. Bony, S., Lu, K.M., & Sud, Y. C. (1997): Sea Surface Temperature and Large-Scale Circulation Influences on Tropical Greenhouse Effect and Cloud Radiative Forcing. *J Of Climate.* Vol. 10 issue 8 (1997)
5. Cano, R., Sordo, C. & Gutiérrez, J.M. (2004): Applications of Bayesian networks in Meteorology. In Gámez, J.A. (Ed.), *Advances in Bayesian networks,* pp. 309-327. Springer-Verlag.
6. Deviren, M. & Khalid, D. (2001): Structural learning of dynamic Bayesian networks in speech recognition. *Proceedings of Eurospeech'2001.* 3-7 Sept. 2001, Aalborg, Denmark . *Retrieved from* http://www.loria.fr/~deviren/publications/MDKD_eurospeech01.pdf
7. Du Plessis, L. A. (2002). A review of effective flood forecasting, warning and response system for applications in South Africa. *Water SA, 28, 129-137.* Retrieved from http://www.wrc.org.za/archives/watersa%20archive/2002/April/1375.pdf
8. Food and Agricultural Organization (FAO), (2004): Drought impact mitigation and prevention in the Limpopo River Basin - A situation analysis. In Corporate Document Repository: *Food and Agriculture Organization of The United Nations, Rome, 2004. Retrieved from website:* http://www.fao.org/docrep/008/y5744e/y5744e00.HTM
9. Famine Early Warning System (FEWS) and United states Agency for International Development (USAID): "Water Requirements Satisfaction Index, (wrsi)". Retrieved from: http://edcintl.cr.usgs.gov/fews/wrsi.html

10. Friedman, N. (1998): "The Bayesian Structural EM algorithm" In *Fourteenth Conference on Uncertainty in Artificial Intelligence, (UAI).*
11. Friedman, N., Murphy, K., & Russel, S. (1998): Learning the structure of dynamic probabilistic networks, In *Proceedings on Uncertainty in Artificial Intelligence, (UAI), p139-147, 1998.*
12. Geiger, D., Heckermean, D., (1995): A characteristic of the Dirichlet distribution with application to learning dynamic Bayesian networks, in *Proceedings of eleventh conference on uncertainty in artificial intelligence.* Morgan Kaufmann Publishers, San Francisco, 196-207, (1995).
13. Ghahramani Z, (1998). Learning dynamic Bayesian networks In C.L. Giles and M. Gori (eds.), Adaptive Processing of Sequences and Data Structures. *Lecture Notes in Artificial Intelligence*, 168-197. Berlin: Springer-Verlag.
14. Heckerman, D. (1999): A Tutorial on Learning with Bayesian networks, in Learning in Graphical Models. *Adaptive computation and machine learning.* The MIT press, Cambridge, Massachusetts, 1st Ed. Retrieved from website: http://research.microsoft.com/research/pubs/view.aspx?msr_tr_id=MSR-TR-95-06
15. Jury, M. (2002): Economic Impacts of Climate Variability in South Africa and Development of Resource Prediction Models. *J of Appl Meteorology, Vol 41, Issue No 1, 2002 p46-55*
16. Kalnay, E. & Coauthors. (1996): The NCEP/NCAR 40 year Reanalysis project. *Bulletin of Amer Met Soc. PP 437-471. Vol. 77, Issue 3, March 1996.*
17. Klein, S. A., Soden, B. J. & Lau, N. (1999): Remote Sea surface Temperature variations during ENSO: Evidence for a Tropical Atmospheric Bridge. *J Of Climate. Vol 12, Issue 4, Apr 1999. P917-932.*
18. Knippertz, P. (2007): Tropical/extra-tropical interactions related to Dynamics of Atmosphere and Oceans. *Dynamics of Atmosphere and Oceans, Vol 43 Issue 1-2 p36-62*
20. Lee, B. & Joseph, J. (2006): Learning a probabilistic model of rainfall using graphical models. *Project reports, School of Computer Science, Carnegie-Mellon University. Machine Learning, 10-701/15-781. Fall 2006.* Retrieved from:http://www.cs.cmu.edu/~epxing/Class/10701-06f/project_reports.html
20. Lu, R. & Dong, B. (2005): Impact of Atlantic sea surface temperature anomalies on the summer climate in the western North Pacific during 1997-

1998. *Journal of Geophysical Research,* Vol. 110, D16102, doi:10.1029/2004JD005676, 2005

21. Makarau, A & Jury Mark, 1997: Seasonal cycle of convective spells over South Africa during Austral Summer. *Int J of Climatology, 17, 1317-1332.*

22. Mason, S and Jury M R, 1997: Climate variability and change over Southern Africa: a reflection on underlying processes. P*rogress in Physical Geography, 21, (23-51)*

23. Matarira C H, 2006: Drought over Zimbabwe in a regional and global context. *Int J of Climatology. Vol 10, Issue 6 pp 609 - 625.*

24. Mihajlovic, V., and Petkovic, M. (2001): dynamic Bayesian networks: A State of the Art. *Technical Report, Centre for Telematics and Information Technology, University of Twente, The Netherlands.* Retrieved from website: http://www.ub.utwente.nl/webdocs/ctit/1/0000006a.pdf

[redacted]Mukabana, J. R. and R.A. Pielke, 1996: Investigating the influence of synoptic scale monsoonal winds and mesoscale circulations on diurnal weather patterns over Kenya using a mesoscale numerical model. *Monthly Weather Review, 124, 224-243.*

26. Murphy K, (1998): A brief introduction to graphic models and Bayesian networks. Retrieved from http://www.cs.ubc.ca/~murphy/Bayes/bnintro.html

27. Murphy K, Mian S, (1999): Modeling gene expression data using dynamic Bayesian networks. *Technical Report, Computer Science Division, University of California, Berkeley, CA.*

28. Murphy, P (2002): dynamic Bayesian network: Representation, Inference and Learning. *PhD Thesis, University Of California, Berkeley.* Available at http://www.cs.ubc.ca/~murphyk/Thesis/thesis.pdf

29. Nicholson, S. E., Leposo, D., & Grist, J. (2001). The relationship between El-Niño and drought over Botswana. *Journal of Climate Volume, 14(3) PP 323-335.*

30. Nicholson, S. E.,& Selato J.C, (2000): The influence of La Nina on African Rainfall. *International Journal of climatology 20: 1761 - 1776.*

31. Neapolitan, R. E. (2004): Learning Bayesian networks. *Prentice Hall Inc,* Upper Saddle River, NJ, USA.

32. Okoola, R. E. (2003). The onset and cessation of the "long rains" in Eastern Africa and their interannual variability. *Theor. Appl. Climatology. 75, 43-54.*

33. Okoola, R. E., 1999. Mid-tropospheric patterns associated with extreme dry and wet episodes over Equatorial Eastern Africa during Northern hemisphere spring. *Journal of Appl. Meteorology. 38, 1161-1169.*

34. Osunmakinde, I. O.,(2009): Evolving dynamic Bayesian networks, Personal Communication from *PhD Research in progress and supervised by Anet Potgieter, Computer Science Department, Faculty of Science, University of Cape Town*

35. Pearl, J. & Verma, T. (1991). A theory of inferred causation, in Principles of Knowledge Representation and Reasoning. *Proceedings of the Second International Conference (Morgan Kaufmann, San Mateo, CA).* Retrieved from http://citeseer.ist.psu.edu/pearl91theory.html

36. Philippon, N. Camberlin, P. & Fauchereau, N. (2002): Empirical predictability study of October - December East African rainfall. *Quarterly J of Royal Met Soc (2002). 128. pp 2239 - 2256.*

37. Preston-Whyte, R. A. (1997): The atmosphere and Weather of Southern Africa. *4th impression, Oxford University Press, Cape Town.*

38. Reason, C.J.C (2002). Sensitivity of the Southern African circulation to dipole sea surface temperature patterns in the South West Indian Ocean. *International Journal of Climatology, **2**, (377-393)*

39. Reason C.J.C. & Mulenga, H. (1999): Relationship between South African rainfall and SST anomalies in the South West Indian Ocean. *International Journal of Climatology **19** (1651-1673)*

40. Reason, C. J. C. & Keibel, A. (2004): Tropical cyclone elain and its unusual penetration and impacts over South African mainland. *Weather and Forecasting. Vol 19. Issue 5, pp 789 - 805 (2004).*

41. Russel, S., & Norvig, P. (2003): Chapter 5 - probabilistic reasoning over time In *Artificial Intelligence: A modern approach*. Prentice Hall, 2nd Edition, 2003.

42. SADC Regional Early Warning Unit, Remote Sensing Unit and USAID FEWSNet. (2005). *Agromet -Update. Regional Food security program.* Issue No. 2 2004-2005.

43. Tadross, M. A. Hewitson, B. C. & Usman, M. T. (2005): The interannual variability of the onset of the maize growing season over South Africa and Zimbabwe. *Journal of climate, 18, 3356-3372.*

44. Todd, M.C., Washington, R., & Palmer, P.I, (2004): Water vapor transport associated with tropical temperature trough system over Southern Africa and South West Indian Ocean. *Int Journal of Climatology, **24**, (555-568).*

45. Umoh, U. T. (2006): Water Resources Management in Botswana under the Increasing Effects of Climate Change, in Proceeding, *Environmentally Sound Technology in Water Resources Management, ACT Press, 2006*

46. Usman, M. T. & Reason, C. J. C. (2004): Dry spell frequencies and their Variability over Southern Africa. *Climate Research, Vol. 26 199-211, 2004*

47. Wang, C. (2002): Atmospheric Circulation Cells Associated with the El-Niño/Southern Oscillation. *J of Climate, Vol. 15. **4**. (2002). p399-419*

48. Washington, R., & Preston, A. (2006): Extreme wet years in southern Africa: the role of the Indian Ocean. *J. Geophys. Res Atmospheres* Vol. 111

49. Washington, R. & Todd, M. (1999): Tropical-temperate links in Southern Africa and South West Indian Ocean satellite derived daily rainfall. *Int J of Climatol. **19**: p1601-1616*

50. World Meteorological Organization (WMO), 2002. Annual Summary of Global Tropical Cyclone Season 2000, (March 2002), *WMO/TD-No.1082. TCP-46*

51. Zukerman, I. & Albrecht, D.W., (2004): Predicting statistical models for user modeling. *School of computer science and software engineering, Monash University, Clayton, Victoria, Australia.* Retrieved from website: http://www.csse.monash.edu.au/~ingrid/Publications/UMUAIiz-dwa.pdf

Appendix 1: Principal Component and Correlation Analysis

The varying nature of Land (albedo, vegetation, soil types, topography, land-use patterns, etc) and Oceans (sea surface temperatures, salinity, Ocean currents, and waves) cause variations in the temporal and spatial characteristics of meteorological observations and therefore it is important to be able to characterize such areas or zones. Zoning is the process of delineating (identifying and separating) a given area of interest into sub-areas which have similar meteorological characteristics (homogeneity) or signal of the meteorological variables.

A homogeneous region can be thought of as an area or region that normally experience similar spatial and temporal characteristics of climatic variables including occurrences of anomalous rainfall, floods or droughts. This similarity could for instance be reflected in the onset and cessation of rains and the generating mechanism or magnitudes of the climate event.

The method of Principal Component Analysis (PCA) has been widely used to delineate rainfall over different countries into regional homogeneous rainfall zones. An example of application of PCA to delineate rainfall zones in East Africa can be obtained from (Basalirwa, Odiyo, Mngodo, & Mpeta, 1999), (Bazira & Ogallo, 1999) and many others. Principal Component Analysis (PCA) which is occasionally referred to as the Empirical Orthogonal Functions (EOF) is derived from Factor Analysis (a statistical technique) used to identify a relatively small number of factors that can be used to represent relationships among sets of many interrelated variables. The basic assumption of factor analysis is that underlying dimensions, or a small number of factors, can be used to explain complex phenomena.

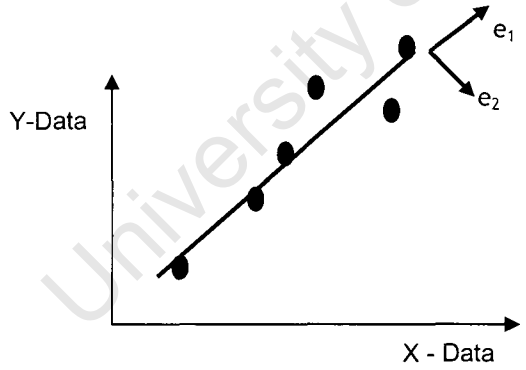
Essentially, PCA consists of a transformation of a large number of related (un-orthogonal) variables into a smaller number of orthogonal (un-related) variables, which present common causes of manifest variable changes.

PCA can therefore be defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. PCA retains those characteristics of the data set that contribute most to its variance, by keeping lower-order principal components and ignoring higher-order ones.

Principles of PCA

- Consider data in station 1: $\{X=(x_1, x2,\ldots\ldots, X_m)\}$ and
  Station 2: $\{= (y_1, y_2,\ldots\ldots, y_m)\}$

- The scatter between X and Y could be as follows:



- e_1 is the vector in the direction of minimum variance of the X-Y data variation and e_2 is orthogonal to e_1 (the direction of minimum variance in this case).
- Thus e_1 and e_2 are orthogonal vectors.
- The X-data can be projected onto the e_1-e_2 axis to give a common value **x** unique for the X-data. In the process a new data U is generated whose dimension is 2 (2 variables in this case).
- U is common to the two variables. As such

$$X = (e_1)^T.U$$

Similarly

$$y = (e_2)^T.U$$

This principal can be extended for many more variables or station than just 2.

- The **U** elements are called *Principle components* (sometimes referred to as principal factors). The $\mathbf{e}^T$ components are derived from the ***orthogonal vectors*** and are called the *component loadings,* (sometimes referred to as *factor loadings).*
- An absolute value of 0.4 for the component loadings is usually adopted to demarcate the extreme boundary of the dominance of the principle component in consideration. A mosaic of the boundaries of all significant orthogonal vectors usually reveals the distribution of the predominant zones of homogeneity.
- There are Mainframe and Personal Computer based software that handles PCA which exist in the market today, e.g. SYSTAT.

Results from Principal Component Analysis for Botswana

	Latent Roots (Eigenvalues)	% Total Variance Explained	Cumulative Variance
1	11.03	38.150	38.150
2	3.977	13.715	51.865
3	1.995	6.878	58.743
4	1.658	5.718	64.461
5	1.571	5.417	69.878
6	1.182	4.077	73.955
7	1.017	3.505	77.460

Modes of PCA

If the parameter under study (e.g. rainfall) is fixed then it is possible to generate a correlation data matrix between various locations (S-Mode) over a set of periods, or between periods (T-Mode) over a set of locations.

The S-mode yields groupings of locations in terms of the *change over time* while the T-mode can yield groupings of periods (Time) with similar *spatial patterns.* In an S-mode analysis, the variables are stations and the observations are the values at each time. The principal component loading matrix will contain the correlation of each station with each component. These can be plotted on a map to depict the spatial pattern of each component. S-mode can be used to classify locations with similar temporal anomalies.

In T-mode analysis, the standardized data matrix is transposed so that each of the individual time periods is changed to a variable while the station names become observations. This analysis produces components with loadings on the individual times. T-mode can be used to classify years in which specific sub-regions experienced similar spatial anomalies.

- To decide on the number of factors to retain for use in representing the data, it is helpful to examine the percentage of the total variance explained by each factor and is usually given by the label Eigenvalue and its corresponding cumulative percentage.
- The proportion of the variance accounted for by the common factors is known as the *communality* and can be obtained by squaring and summing up the corresponding significant factor loadings.

- One should apply their knowledge about topography and other underlying features affecting the rainfall distribution of their country to come up with realistic groupings.

- The final homogeneous regional groupings should be a product (combination) of the contribution of the spatial patterns using all the significant factor loadings after taking into consideration the vector plot clusters and inter-station correlation coefficients.

- The Figures below are homogeneous regional grouping for Botswana

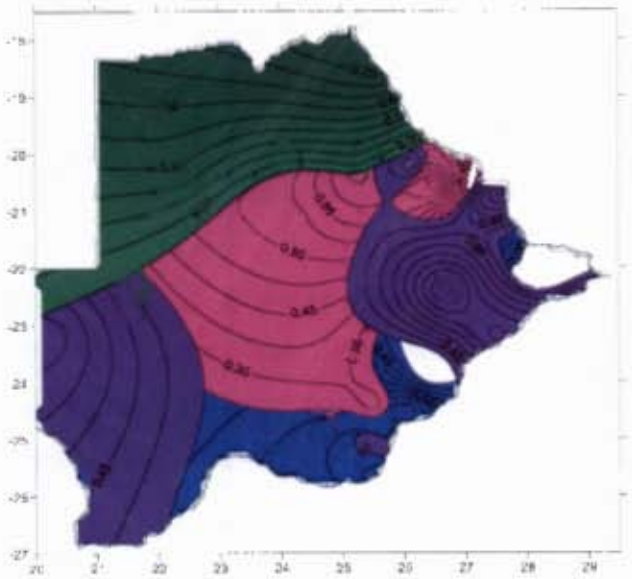
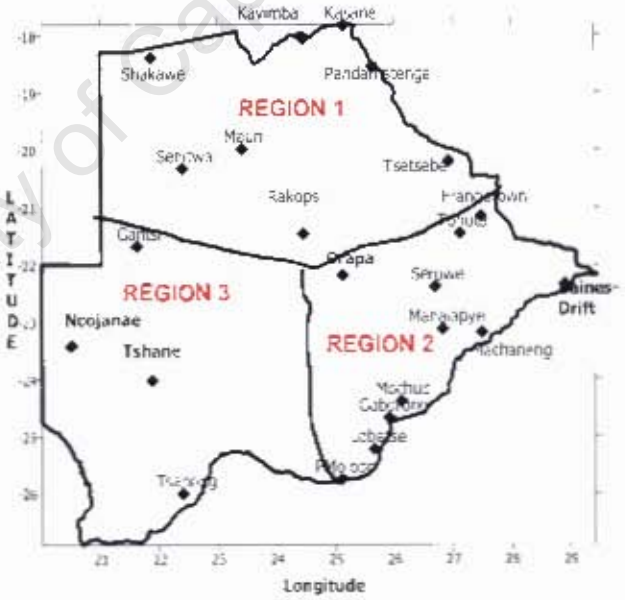
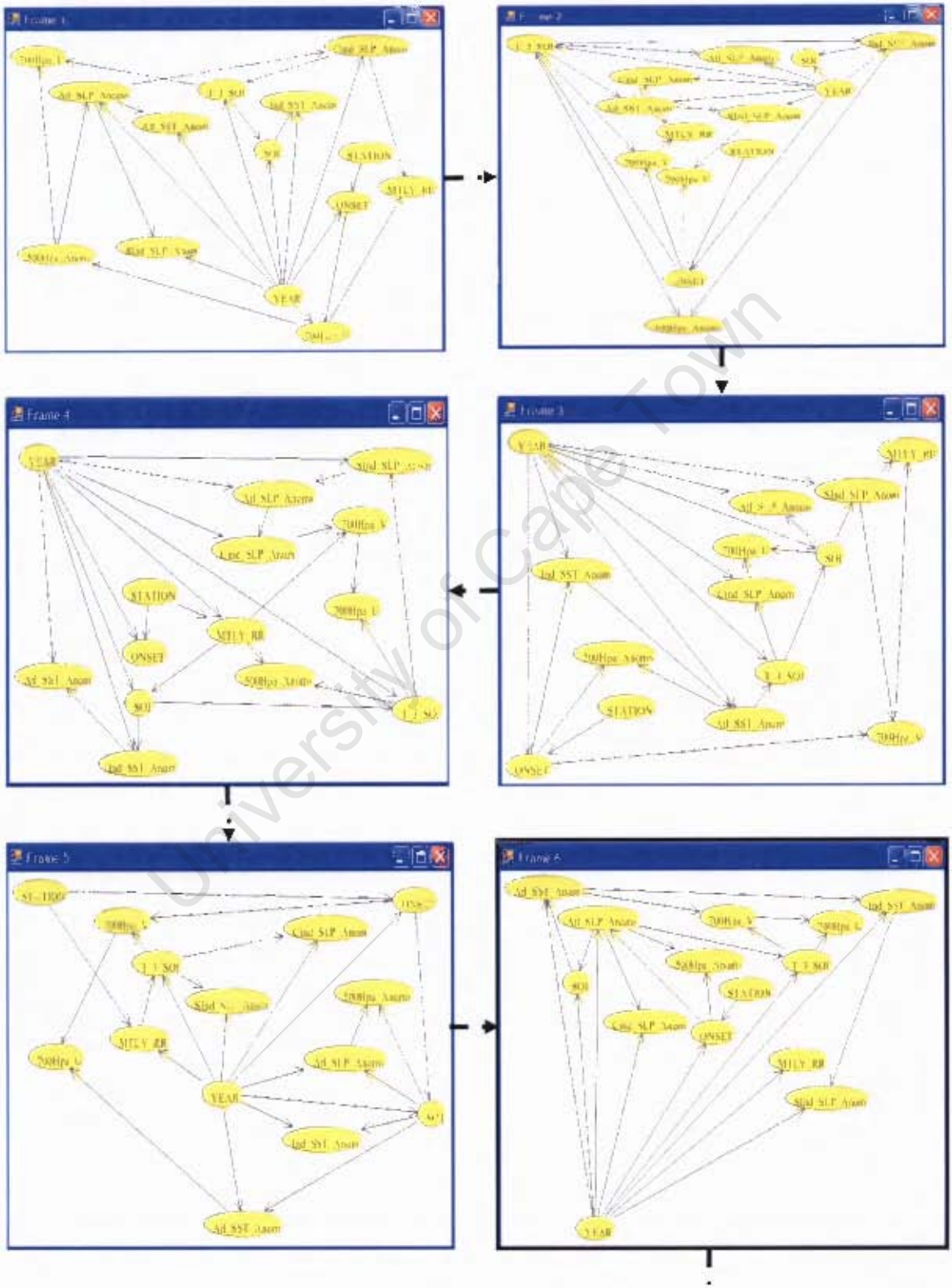


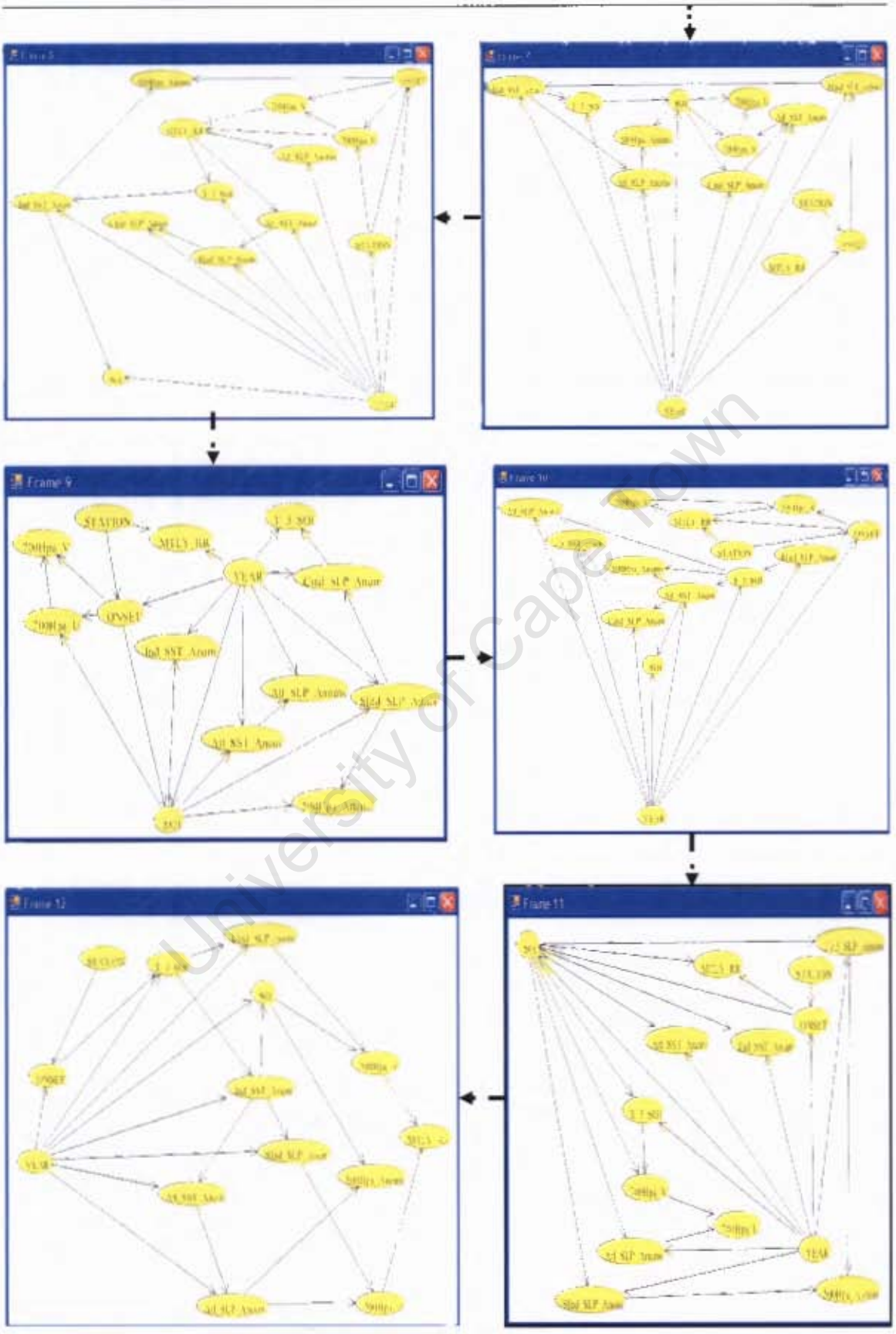
Figure 38.Preliminary and final plots of homogeneous zones over Botswana



Appendix 3: DBN representing January to December months

Figure 39: Complete set of Frames 1-12 of DBN representing January to December months.





Appendix 4: Sample of data organized for analysis with the ESA

94	2001	10	48.5	0.457	0.470	-1.9	-3	-0.115	0.848	0.140	9.494	3.093	-1.469	2
94	2001	11	57.3	-0.044	0.272	7.2	-8.9	0.507	1.213	0.969	4.759	1.581	-1.560	2
94	2001	12	75.0	-0.010	0.767	-9.1	1.4	0.494	-0.780	1.388	7.815	2.610	-0.443	2
94	2002	1	25.0	0.213	0.763	2.7	-1.9	0.437	1.369	-0.730	8.152	1.132	-1.828	2
94	2002	2	30.0	0.152	0.415	7.7	7.2	0.632	0.551	1.564	16.898	0.850	-0.104	2
94	2002	3	0.0	0.014	0.295	-5.2	-9.1	-0.178	0.577	-0.304	16.916	-0.448	-0.018	2
94	2002	4	0.0	-0.306	0.558	-3.8	2.7	-2.041	1.285	-0.207	8.623	0.276	-1.415	2
94	2002	5	0.0	-0.186	0.443	-14.5	7.7	0.873	0.152	-0.391	10.814	3.572	-0.691	2
94	2002	6	0.0	0.074	0.424	-6.3	-5.2	2.475	1.498	0.350	4.776	3.582	-1.060	2
94	2002	7	0.0	-0.013	0.824	-7.6	-3.8	0.087	-1.427	1.262	-1.862	7.564	-0.120	2
94	2002	8	0.0	0.086	0.431	-14.6	-14.5	0.896	2.213	1.181	15.511	4.364	-3.189	2
94	2002	9	0.0	-0.028	0.529	-7.6	-6.3	0.896	2.213	1.181	15.511	-999.900	-999.900	2

STATION	YEAR	MONTH	MTLY_RR	Atl_Anom	Ind_Anom	SOI	(T-3)SOI	SLP_Ano	Ind_SLP_Anom	d_SLP_Ano	Hpa_Anoi	700Hpa_U	700Hpa_V	ONSET
251	1971	1	107.2	-0.081	-0.329	0.3	0.9	-1.303	-1.571	-1.400	-14.047	-0.418	0.898	999
251	1971	2	22.5	-0.150	-0.630	1.9	1.7	-1.308	-1.789	-2.966	-18.126	-0.496	0.936	999
251	1971	3	19.9	-0.286	-0.498	2.1	2.1	-0.728	-0.153	0.716	3.759	0.768	-1.342	999
251	1971	4	10.2	-0.099	-0.358	1.7	0.3	-0.221	-2.295	-0.397	6.319	3.647	-1.038	999
251	1971	5	15.6	0.096	-0.133	0.7	1.9	-0.467	-0.558	-0.361	-16.870	3.296	-1.766	999
251	1971	6	0	-0.143	-0.181	0.1	2.1	-0.395	1.278	0.190	-8.452	5.161	-1.197	999
251	1971	7	0	-0.163	-0.224	0.1	1.7	-0.623	-1.377	-0.558	-20.335	6.334	-2.975	1
251	1971	8	0	-0.261	-0.364	1.3	0.7	-0.254	-2.747	-0.839	-22.333	8.720	-3.710	1
251	1971	9	0	-0.350	-0.314	1.6	0.1	-0.154	-1.028	-0.285	3.947	6.549	-3.164	1
251	1971	10	5.3	-0.320	-0.296	1.7	0.1	0.265	0.368	0.960	-3.010	5.764	-2.292	1
251	1971	11	16.9	-0.577	-0.371	0.5	1.3	0.477	-0.567	-0.351	-20.148	3.580	-1.658	1
251	1971	12	26.3	-0.715	-0.407	0	1.6	-0.346	-1.090	-0.352	-11.048	2.628	-0.231	1
251	1972	1	190.5	-0.438	-0.310	0.4	1.7	-1.773	-0.811	-0.240	-15.747	-1.696	0.349	1
251	1972	2	10.7	-0.219	-0.269	0.8	0.5	-0.488	-3.059	-1.416	-20.356	-0.324	3.282	1
251	1972	3	109.3	-0.576	-0.263	0.1	0	-0.818	0.127	0.867	-11.741	-0.020	-1.149	1
179	1991	1	153.0	-0.091	0.267	5.1	1.8	1.327	1.519	0.590	14.4326	-3.872	-1.580	4
179	1991	2	115.0	-0.147	0.161	0.6	-5.3	1.072	-1.289	0.614	12.0781	-2.495	1.430	4
179	1991	3	47.5	-0.059	0.337	-10.6	-2.4	-0.068	0.337	0.696	4.73633	-3.340	-0.112	4
179	1991	4	0.0	-0.171	0.208	-12.9	5.1	0.079	1.875	-0.497	14.0425	-4.087	0.996	4
179	1991	5	0.0	-0.220	0.449	-19.3	0.6	-0.417	-0.138	-0.851	0.603516	-1.537	1.166	4
179	1991	6	0.0	0.143	0.598	-5.5	-10.6	0.735	-0.962	-1.200	-20.8833	2.870	-1.918	4
179	1991	7	0.0	0.019	0.526	-1.7	-12.9	0.847	0.073	0.462	10.4482	-0.770	-2.064	3
179	1991	8	0.0	-0.024	0.373	-7.6	-19.3	1.936	-0.617	-0.109	24.3311	2.141	-0.894	3
179	1991	9	0.0	0.009	0.141	-16.6	-5.5	0.626	-1.028	0.735	12.9668	2.009	-2.308	3
179	1991	10	13.0	-0.316	0.219	-12.9	-1.7	0.735	-0.472	-2.160	-0.63623	-1.091	-1.754	3
179	1991	11	28.0	-0.315	0.215	-7.3	-7.6	0.867	1.093	-0.031	9.26953	-2.958	0.268	3
179	1991	12	37.0	-0.257	0.127	-16.7	-16.6	-0.526	1.830	-0.782	0.634766	-3.510	0.042	3
179	1992	1	42.0	-0.133	0.216	-25.4	-12.9	0.917	1.149	1.340	21.6123	-3.177	1.194	3
179	1992	2	12.5	-0.041	0.237	-9.3	-7.3	1.102	2.621	-1.526	30.8984	-2.664	1.298	3
179	1992	3	41.0	-0.161	-0.013	-24.2	-16.7	0.942	-0.753	0.247	8.95605	-1.940	-0.048	3
179	1992	4	7.6	-0.477	0.212	-18.7	-25.4	0.669	0.535	0.763	10.5127	-2.766	0.494	3
179	1992	5	0.0	-0.555	0.136	0.5	-9.3	1.113	-0.828	0.159	8.37354	0.331	0.598	3
179	1992	6	1.5	-0.587	-0.086	-12.8	-24.2	0.685	-0.862	-0.240	8.30664	-0.166	-0.184	3
179	1992	7	0.0	-0.523	0.115	-6.9	-18.7	1.267	0.553	0.822	20.5181	-2.030	-2.366	2
179	1992	8	0.0	-0.470	-0.090	1.4	0.5	1.966	3.043	1.051	3.29102	0.929	-2.508	2
179	1992	9	0.0	-0.373	-0.453	0.8	-12.8	0.406	0.662	0.235	7.1167	-1.505	-2.322	2
179	1992	10	3.5	-0.387	-0.400	-17.2	-6.9	0.535	-0.542	0.190	4.25391	-0.570	-3.068	2
179	1992	11	50.5	-0.515	-0.059	-7.3	1.4	0.557	-0.227	0.289	2.0293	-0.423	-1.200	2
179	1992	12	40.5	-0.266	-0.294	-5.5	0.8	0.514	-0.510	1.838	-0.78516	-3.963	0.548	2
179	1993	1	101.5	-0.282	-0.272	-8.2	-17.2	0.167	-1.761	0.790	10.7222	-3.183	1.946	2
179	1993	2	128.5	-0.247	-0.020	-7.9	-7.3	0.242	0.781	1.664	1.30811	-4.769	-0.844	2
179	1993	3	7.5	-0.297	0.001	-8.5	-5.5	-0.048	-0.103	0.326	0.226074	-3.870	2.338	2
179	1993	4	1.0	0.069	-0.077	-21.1	-8.2	0.629	0.635	1.253	4.11279	0.060	-1.262	2
179	1993	5	0.0	0.214	-0.064	-8.2	-7.9	-0.097	-2.438	-0.011	11.314	2.426	-0.128	2
179	1993	6	0.0	0.125	-0.234	-16	-8.5	0.855	1.908	0.610	24.7568	0.425	-0.018	2
179	1993	7	2.0	0.054	-0.210	-10.8	-21.1	-0.683	3.363	1.272	15.8281	-2.989	-1.528	2
179	1993	8	0.0	-0.007	0.026	-14	-8.2	-0.224	1.883	0.901	2.01123	0.249	-1.074	2
179	1993	9	19.0	0.205	-0.045	-7.6	-16	-1.384	2.282	1.025	8.23682	-3.275	0.622	2
179	1993	10	15.0	0.200	-0.091	-13.5	-10.8	-0.455	2.078	1.220	22.4736	-1.093	-0.228	2
179	1993	11	28.0	0.219	-0.062	0.6	-14	-0.863	1.133	-0.391	-2.02051	-1.069	-1.232	2
179	1993	12	81.0	0.186	-0.347	1.6	-7.6	0.514	1.230	0.838	21.6748	-2.555	0.016	2