UNIVERSITEIT•STELLENBOSCH•UNIVERSITY
jou kennisvennoot • your knowledge partner

SKRIPSIE TITLE ¡TODO¿

Coen Potgieter
25999656

¡TODO: This is probably a place holder, not sure what to put here though...¿
Report submitted in partial fulfilment of the requirements of the module Project (E) 448 for the degree Baccalaureus in Engineering in the Department of Electrical and Electronic Engineering at Stellenbosch University.

Supervisor: Dr C. Van Daalen
¡TODO: ASK IF THIS IS FINE¿

November 2025

Acknowledgements

¡TODO: Do Really need this?¿

I would like to thank my dog, Muffin. I also would like to thank the inventor of the incubator; without him/her, I would not be here. Finally, I would like to thank Dr Herman Kamper for this amazing report template.

UNIVERSITEIT•STELLENBOSCH•UNIVERSITY
jou kennisvennoot • your knowledge partner

Plagiaatverklaring / *Plagiarism Declaration*

1. Plagiaat is die oorneem en gebruik van die idees, materiaal en ander intellektuele eiendom van ander persone asof dit jou eie werk is.
   Plagiarism is the use of ideas, material and other intellectual property of another's work and to present is as my own.

2. Ek erken dat die pleeg van plagiaat 'n strafbare oortreding is aangesien dit 'n vorm van diefstal is.
   I agree that plagiarism is a punishable offence because it constitutes theft.

3. Ek verstaan ook dat direkte vertalings plagiaat is.
   I also understand that direct translations are plagiarism.

4. Dienooreenkomstig is alle aanhalings en bydraes vanuit enige bron (ingesluit die internet) volledig verwys (erken). Ek erken dat die woordelikse aanhaal van teks sonder aanhalingstekens (selfs al word die bron volledig erken) plagiaat is.
   Accordingly all quotations and contributions from any source whatsoever (including the internet) have been cited fully. I understand that the reproduction of text without quotation marks (even when the source is cited) is plagiarism

5. Ek verklaar dat die werk in hierdie skryfstuk vervat, behalwe waar anders aangedui, my eie oorspronklike werk is en dat ek dit nie vantevore in die geheel of gedeeltelik ingehandig het vir bepunting in hierdie module/werkstuk of 'n ander module/werkstuk nie.
   I declare that the work contained in this assignment, except where otherwise stated, is my original work and that I have not previously (in its entirety or in part) submitted it for grading in this module/assignment or another module/assignment.

Studentenommer / *Student number*	Handtekening / *Signature*
Voorletters en van / *Initials and surname*	Datum / *Date*

Abstract

English

The English abstract.

Afrikaans

Die Afrikaanse uittreksel.

Contents

Declaration ii

Abstract iii

List of Figures vi

List of Tables vii

Nomenclature viii

1. Introduction 1

1.1. Background . . . 1

1.1.1. Drought as a growing threat . . . 1

1.1.2. Water demand, vulnerability, and regional impact . . . 1

1.1.3. Complexity Of Drought . . . 2

1.1.4. Towards integrated drought monitoring in South Africa . . . 4

1.2. Problem Statement . . . 5

1.3. Project Objectives . . . 5

1.4. Summary Of Work . . . 5

1.5. Scope . . . 5

1.6. Roadmap . . . 6

1.7. THIS IS THEIR TUT . . . 7

1.8. Section heading . . . 7

2. Literature Review 9

3. Methods 14

3.1. Data Acquisition . . . 14

3.1.1. Sources . . . 14

3.1.2. Preprocessing . . . 14

3.2. Index Calculation . . . 15

3.2.1. SPI . . . 15

3.2.2. SDI . . . 15

3.2.3. NDVI . . . 15

3.2.4. Discretisation of Indices . . . 16

3.3. Model Development 16
3.3.1. Model Design 16
3.3.2. Inference 20
3.3.3. Parameter Estimation 23
3.3.4. Model Selection 26
3.4. Model Implementation 28
3.4.1. Programming Environment & Tools 28
3.4.2. Data Pipeline Implementation 28
3.4.3. Model Implementation 30
3.4.4. Model Selection & Output 31

4. Results **32**
4.1. Model Decoding 32
4.1.1. Viterbi 32
4.1.2. MPM Rule 32

5. Body **33**
5.1. Model Selection 33
5.2. Meditating a little bit more on model output 33
5.2.1. What I have been doing 33
5.3. Viterbi Algorithm 33

6. Summary and Conclusion **35**

Bibliography **36**

A. Project Planning Schedule **41**

B. Outcomes Compliance **42**

List of Figures

1.1. I am the short caption that appears in the list of figures, without references. 8

3.1. TODO 17
3.2. Junction Tree representation of the DNBC. Each cluster groups together latent state variables and observed attributes, with sepsets defined along the edges. Messages are propagated through the tree to perform exact inference. 20
3.3. Data pipeline for DNBC inputs 29

List of Tables

1.1. Performance of the unconstrained segmental Bayesian model on TIDigits1 over iterations in which the reference set is refined. 7
1.2. A table with an example of using multiple columns. 7

3.1. Discretisation thresholds for drought indices. 16
3.2. Summary of random variables in the model 17
3.3. Priors Factor Table 18
3.4. Transition Factor Table & Transition Matrix 19
3.5. Emission Factor Table 19
3.6. Cluster potentials for the DNBC. Each potential corresponds either to a state transition or to a state-attribute relationship. 21

Nomenclature

Variables and functions

S_t	TODO
$A_t^{(n)}$	TODO
$a_t^{(n)}$	TODO
$a_{i,j}$	The probability of a transition from HMM state s_i to state s_j.
$b_i^{(n)}(j)$	TODO.
$p(x)$	Probability density function with respect to variable x.
$P(A)$	Probability of event A occurring.
ε	The Bayes error.
ε_u	The Bhattacharyya bound.
B	The Bhattacharyya distance.
s	An HMM state. A subscript is used to refer to a particular state, e.g. s_i refers to the i^{th} state of an HMM.
$\mathbf{S}$	A set of HMM states.
$\mathbf{F}$	A set of frames.
$\mathbf{o}_f$	Observation (feature) vector associated with frame f.
$\gamma_s(\mathbf{o}_f)$	A posteriori probability of the observation vector $\mathbf{o}_f$ being generated by HMM state s.
μ	Statistical mean vector.
Σ	Statistical covariance matrix.
$L(\mathbf{S})$	Log likelihood of the set of HMM states $\mathbf{S}$ generating the training set observation vectors assigned to the states in that set.
$\mathcal{N}(\mathbf{x}\|\mu, \Sigma)$	Multivariate Gaussian PDF with mean μ and covariance matrix Σ.
N	Total number of frames or number of tokens, depending on the context.
D	Number of deletion errors.
I	Number of insertion errors.
S	Number of substitution errors.

Acronyms and abbreviations

DNBC	Dynamic Naive Bayes Classifier
AE	Afrikaans English
AID	accent identification
ASR	automatic speech recognition
AST	African Speech Technology
CE	Cape Flats English
DCD	dialect-context-dependent
DNN	deep neural network
G2P	grapheme-to-phoneme
GMM	Gaussian mixture model
HMM	hidden Markov model
HTK	Hidden Markov Model Toolkit
IE	Indian South African English
IPA	International Phonetic Alphabet
LM	language model
LMS	language model scaling factor
MFCC	Mel-frequency cepstral coefficient
MLLR	maximum likelihood linear regression
OOV	out-of-vocabulary
PD	pronunciation dictionary
PDF	probability density function
SAE	South African English
SAMPA	Speech Assessment Methods Phonetic Alphabet

Chapter 1

Introduction

1.1. Background

1.1.1. Drought as a growing threat

Climate change is no longer a distant projection, but rather, is already reshaping how frequently and how severely extreme weather events occur. Recent reports and studies indicate that the frequency and intensity of droughts have markedly increased worldwide since the early 21st century. For instance, the OECD's Global Drought Outlook reports that approximately 40% of global land experienced upticks in both drought frequency and intensity when comparing the periods 1950-2000 to 2000-2020 [1]. Nature's "Warming accelerates global drought severity" highlights that, globally, drought magnitude has become more negative and that the number of drought months is increasing under observed climate conditions, whilst it is also being reported that multiyear droughts are becoming increasingly common [2, 3].

1.1.2. Water demand, vulnerability, and regional impact

As global population continues to climb in South Africa at a rapid rate, water demand increases. Agriculture, industry and urban use all place stress on water systems, of which are already under threat. Poor infrastructure and inequitable management are two prevalent issues in the nation which only exacerbate the cost of drought [2, 4]. Africa has been particularly vulnerable: since the 1960s, more than 382 drought events have affected millions of people, especially in Sahel and Southern Africa [1, 5]. In South Africa, severe droughts have left lasting socio-economic scars: notable events include 1973-74, 1983-84, 1991-92, 1994-95, 2014-16, and 2017-18, each associated with sharp losses in crop yields, dam storages, and human hardship [6–9].

The severe 1981–1984, multi-year drought across southern Africa demonstrated that water deficits in the region can be persistent and continent-scale. Recent climate analyses characterise the early 1980s event as among the most pronounced multi-annual rainfall deficits in the twentieth century for southern Africa. Consequences included widespread crop and livestock losses, major food-security interventions and sustained

economic hardships in rural livelihoods that, in some catchments, persisted for several years after precipitation recovered. Such historical events are important because they illustrate not only acute system stress but also the long tail of socio-economic recovery following protracted drought.

A second, and more recent episode is the 2015–2018 drought in the Western Cape which revealed multiple systemic vulnerabilities in both infrastructure and governance. The region experienced severe municipal restrictions as reservoir storages declined to between roughly 15–30% of capacity, provoking near-municipal "Day Zero" scenarios, emergency demand management and extraordinary conservation measures. The drought also produced substantial agricultural economic losses, associated labour reductions, and marked pressures on public-health and social services [7, 9, 10].

The crisis in the Western Cape also exposed the limits of urban water supply designs that assume relatively steady inter-annual availability, and it highlighted institutional gaps in reservoir operation, intergovernmental coordination and demand-side planning. Analyses of the City of Cape Town response emphasise how communications, behavioural change and temporary policy levers averted the most catastrophic outcomes, but also that these were last-resort measures that imposed disproportionate burdens on low-income communities and agricultural producers dependent on the urban market. Reports and post-event reviews point to the need for improved system modelling, diversified supply portfolios and explicit drought contingency plans at municipal and provincial levels [11, 12].

Drought has direct consequences for agricultural productivity, human and animal health, and vegetation cover, with water scarcity leading to food insecurity and poverty [5]. Indirectly, drought can contribute to environmental degradation, exacerbate food shortages, diminish human welfare, and, in certain contexts, act as a catalyst for social unrest [13]. Across Africa, the agricultural sector has borne significant impacts, manifesting as the degradation of grazing lands, crop failure, depletion of farming assets, and the impoverishment of farmers, particularly vulnerable smallholder farmers, often culminating in forced migration from rural to urban areas [5].

South Africa's recent and historical droughts make clear that water scarcity is a clear risk that is worsened by poor infrastructure, governance constraints and socio-economic inequality. This points to the need for more integrated monitoring and decision-support tools.

1.1.3. Complexity Of Drought

Not only are the impacts of drought multifaceted, but drought itself is a complex and multifaceted phenomenon that resists a simple or universal definition [14]. Unlike discrete natural disasters such as floods or earthquakes, drought unfolds gradually, often with indistinct onset and termination periods. This complexity arises from the fact that drought

is not merely a physical phenomenon but a convergence of meteorological, hydrological, agricultural, and socio-economic processes, as defined by Wilhite and Glantz [15]. Consequently, researchers and policymakers have approached the study and monitoring of drought through a wide range of indices, each of which seeks to capture one particular dimension of this broader phenomenon.

Let us now look at a brief explanation of each category. Meteorological drought is defined as a period of significantly below-average precipitation, which typically serves as the primary trigger for drought conditions and is often quantified by indices such as the Standardised Precipitation Index (SPI). This index measures how much precipitation deviates from the long-term average, normalized to a standard normal distribution. It can be computed for different time scales, most commonly for 1-month, 3-month, 6-month & 12-month [16, 17].

However, such meteorological measures alone cannot capture subsequent and cumulative effects on hydrological systems, ecosystems and human livelihoods. Hydrological drought describes reductions in surface and subsurface water resources, such as streamflow, groundwater tables, reservoir storage, etc. This type of drought is typically lagged behind meteorological drought; it is measured using indices such as the Streamflow Drought Index (SDI) and metrics derived from river monitoring [18, 19].

Agricultural drought describes the phenomenon where the climate interacts with the agriculture to cause a significant decline in production or a deterioration in crop yield and/or quality. Consequently, its measurement focuses on soil moisture availability, crop yield, and vegetation health. The latter is increasingly quantified using remote sensing indices like the Normalized Difference Vegetation Index (NDVI) [20]. It is important to note that agricultural drought is a broader concept than purely meteorological drought, as it can be induced or exacerbated by non-environmental factors. However, these socio-economic factors, such as inadequate irrigation infrastructure or poor land management practices, often determine the severity of the impact that a precipitation deficit has on agricultural output [21].

Socio-economic drought encompasses the human consequences of water scarcity and agricultural failure: it occurs when demand for water, food or energy exceeds supply due to drought disruptions, manifesting in outcomes such as food insecurity, income loss, migration or social unrest [22]. Although socio-economic drought is difficult to quantify directly, researchers have attempted to capture it via composite indices integrating the three types of drought mentioned above and/or by applying vulnerability and economic or social indicators to measure human exposure and impacts [23, 24].

To make matters worse, these different facets of drought manifest differently across South Africa's varying climate zones. The Western Cape sees winter-rainfall with a Mediterranean climates, the east coast sees summer-rainfall and subtropical climates, while the interior regions of the country are semi-arid. This spatial heterogeneity alters the

timing, lag and propagation of drought [25, 26].

Indices designed for a single disciplinary perspective (meteorological, hydrological or agricultural) will emphasise different events and different timings. This will produce diverging or noisy signals that complicate interpretation, leading to poor decision making. In a country with contrasting rainfall regimes this means that a single index cannot reliably capture exposure, vulnerability and impact across all regions. This is a core reason to pursue integrated or composite monitoring approaches [27].

1.1.4. Towards integrated drought monitoring in South Africa

Conventional drought indices each capture a particular physical or ecological dimension of drought. Namely, the SPI for meteorological, SDI for hydrological, and NDVI for agricultural or ecological stress. Relying on any single index therefore provides an incomplete view. Often times these indices contradict each other and show substantial noise requiring an industry experts to diligently analyse them, ultimately leading to false positives and negatives for different users and complicates decision-making when policymakers require a consistent, interpretable drought declaration [28].

A composite indicator aims to combine the output of different, well-established indices to gain a more holistic assessment of drought exposure and its impacts. The benefits include improved detection of drought impacts, more robust signals through redundancy across inputs, and clearer communication to stakeholders who require an integrated risk of drought. Composite models such as the U.S. Drought Monitor and the European Combined Drought Indicator demonstrate how convergent evidence can be used to perform weekly or monthly monitoring. It should be noted that composite approaches are not plug-and-play; they require careful design choices and are sensitive to input quality [29–31].

South Africa has made progress in index development and in the use of multiple indices, but the literature and operational practice still lack a widely-adopted, national composite drought product akin to the USDM or the EDO-CDI mentioned above. Recent reviews of drought monitoring in southern Africa highlight that integrated, multivariate approaches are increasingly recommended, however, composite indices in a South African context remain scarce [25, 32].

There are also existing studies that motivate this project. Dynamic Naive Bayes Classifiers (DNBC) have been recently used successfully in other countries, most notably in South Korea, to combine individual indices into an integrated multiple-drought index. These studies showed improved detection through the output of their probabilistic models compared with single indices alone. They illustrate the technical feasibility of the DNBC approach and provide a methodological blueprint for adapting such a classifier to a South African context. Crucially, however, the transfer of these methods to South Africa requires careful calibration to local climates, and of course, data availability [33, 34].

1.2. Problem Statement

South Africa lacks an operational composite drought indicator that integrates meteorological, hydrological and agricultural dimensions. This project addresses that gap by developing and evaluating a Dynamic Naive Bayes Classifier that combines SPI, SDI and NDVI for drought monitoring in a principled approach.

1.3. Project Objectives

The overarching aim of this study is to advance drought monitoring in South Africa by developing and testing an integrated, probabilistic approach. To this end, three specific objectives were pursued:

1. Develop a composite drought indicator using a DNBC, designed to integrate meteorological, hydrological and agricultural dimensions of drought through the SPI, SDI and NDVI.

2. Assess the performance of this DNBC-based indicator against each of the indices individually. Thus, one can gauge whether the composite framework provides improved, or even comparable, detection of drought events.

3. Finally, Apply the model to the South African context, to evaluate the applicability of this approach.

Together, these objectives define the scope of the study and provide clear criteria against which the success of the project is evaluated.

1.4. Summary Of Work

- Here we tell the reader what was achieved. The idea is for the marker to see what was done here and the rest of the report is proof of this.
- Not sure how this is different from the problem Statement, but okay.
- Maybe go into more detail

1.5. Scope

- This section is for what you didnt do.
- The idea is to show the reader that you considered the problem properly

- Mention this is a positive light: I didn't do this because of x and x, but rather did this

1.6. Roadmap

Simply explain what you did in each chapter

1.7. THIS IS THEIR TUT

The last few years have seen great advances in speech recognition. Much of this progress is due to the resurgence of neural networks; most speech systems now rely on deep neural networks (DNNs) with millions of parameters [**?**]. However, as the complexity of these models has grown, so has their reliance on labelled training data. Currently, system development requires large corpora of transcribed speech audio data, texts for language modelling, and pronunciation dictionaries. Despite speech applications becoming available in more languages, it is hard to imagine that resource collection at the required scale would be possible for all 7000 languages spoken in the world today.

I really like apples.

1.8. Section heading

This is some section with two table in it: Table 1.1 and Table 1.2.

Table 1.1: Performance of the unconstrained segmental Bayesian model on TIDigits1 over iterations in which the reference set is refined.

Metric	1	2	3	4	5
WER (%)	35.4	23.5	21.5	21.2	22.9
Average cluster purity (%)	86.5	89.7	89.2	88.5	86.6
Word boundary F-score (%)	70.6	72.2	71.8	70.9	69.4
Clusters covering 90% of data	20	13	13	13	13

Table 1.2: A table with an example of using multiple columns.

	Accuracy (%)		
Model	Intermediate	Output	Bitrate
Baseline	27.5	26.4	116
VQ-VAE	26.0	22.1	190
CatVAE	28.7	24.3	215

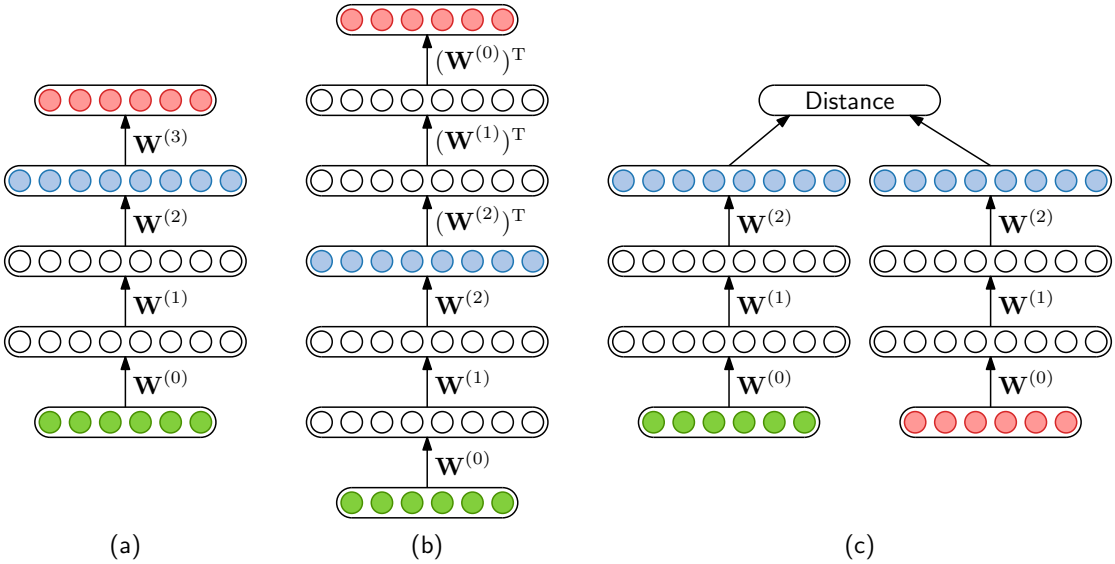

Figure 1.1: (a) The cAE as used in this chapter. The encoding layer (blue) is chosen based on performance on a development set. (b) The cAE with symmetrical tied weights. The encoding from the middle layer (blue) is always used. (c) The siamese DNN. The cosine distance between aligned frames (green and red) is either minimized or maximized depending on whether the frames belong to the same (discovered) word or not. A cAE can be seen as a type of

This is a new page, showing what the page headings looks like, and showing how to refer to a figure like Figure 1.1.

The following is an example of an equation:

$$P(\mathbf{z}|\boldsymbol{\alpha}) = \int_{\boldsymbol{\pi}} P(\mathbf{z}|\boldsymbol{\pi})\, p(\boldsymbol{\pi}|\boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\pi} = \int_{\boldsymbol{\pi}} \prod_{k=1}^{K} \pi_k^{N_k} \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}\, \mathrm{d}\boldsymbol{\pi} \tag{1.1}$$

which you can subsequently refer to as (1.1) or Equation 1.1. But make sure to consistently use the one or the other (and not mix the two ways of referring to equations).

Chapter 2

Literature Review

Introduction

An attempt at advancing drought monitoring depends on a substantial foundation of prior research. Scholars have investigated methods ranging from traditional single-index approaches to more sophisticated probabilistic models. Notably, South Korean research has successfully employed Dynamic Naïve Bayes Classifiers (DNBCs) to develop composite drought indicators, and Hidden Markov Models (HMMs) have been used elsewhere to model drought dynamics. In contrast, the majority of South African studies concentrate on single indices for regional analyses, paying little attention to composite indicators. This review synthesises key contributions from these research areas to provide necessary context before continuing.

Development of a Multiple-Drought Index for Comprehensive Drought Risk Assessment Using a Dynamic Naive Bayesian Classifier

In this study the authors developed a Dynamic Naive Bayesian Classifier multiple-drought index (DNBC-MDI) to produce a probabilistic, multi-dimensional assessment of drought risk. Their stated objectives were to combine conventional drought indices (SPI, SDI, ESI and WSCI) using a DNBC, to apply the resulting DNBC-MDI to bivariate drought-frequency analysis for risk estimation, and to investigate future changes in drought risk under an RCP8.5 climate scenario. Methodologically, the study focused on the Han River basin and used observed data for 1974–2016 together with synthetic climate projections for 2017–2099 generated by the HadGEM2-AO model under RCP8.5 scenario. The DNBC parameters were estimated with an expectation–maximisation (EM) algorithm using the `depmixS4` package from R. Bivariate drought frequency was assessed using a Clayton copula, and a risk equation was employed to compute 100-year return-period risks. The principal results showed that the DNBC-MDI achieved the highest average classification accuracy compared with the individual indices, whilst successfully reproducing several known drought episodes (1994–1995, 2001, 2008–2009, 2012, 2014–2015). The authors were very candid about their limitations, which are as follows:

1. The assessment focused predominantly on climate model outputs, disregarding

remote-sensing products. For this they suggest using MODIS or Landsat.

2. The model assumes conditional independence among the input indices. This assumption is brittle given the interconnections between precipitation, streamflow and evapotranspiration processes.

Overall, the paper demonstrates the technical feasibility and potential advantages of a DNBC-based composite indicator for drought characterisation. Simultaneously, it also signals important areas of concern with regards to robustness and transferability for an adaptation [33].

Assessment of Probabilistic Multi-Index Drought Using a Dynamic Naive Bayesian Classifier

This paper wanted to apply a DNBC to integrate multiple drought indices into a single, coherent drought state representation. The objectives were to combine indicators from different feature spaces, that being: SPI for meteorological, SDI for hydrological, and NVSWI for agricultural. Additionally, they wanted to evaluate whether the DNBC-based drought states could outperform individual indices in terms of detection, classification, and persistence. The study was carried out in the Han River upstream sub-basin in South Korea, using data from 1980–2015 for in-situ observations and 2003–2015 for MODIS-derived indices. The DNBC was constructed with five hidden drought states, the number selected using AIC, BIC and minimum log-likelihood for model selection criteria, and parameters were estimated through the EM algorithm implemented in the `depmixS4` R package.

The key results showed that the DNBC-based drought states successfully reproduced known drought episodes (2004, 2006, 2008–2009, 2014, 2015) and provided accurate representations of drought duration and persistence. In detection performance, DNBC-DS captured 100%, 96%, 100%, and 93% of droughts identified by SPI, SDI, NVSWI, and a composite drought index (CDI) respectively. The approach also highlighted the differing relationships between indicators, with strong correlation between SPI and SDI with a score of 0.648, but weak correlations involving NVSWI (0.186–0.187). Overall, the DNBC offered a probabilistic framework for drought monitoring that explicitly incorporated uncertainty, outperforming deterministic single-index approaches.

Regardless, the authors acknowledged some of their key limitations. Firstly, the model relied on only three indices, which is not complex enough to capture what we call drought. It excluded potentially informative variables such as temperature, water vapour, and radiation. Beyond these, aligning with the paper above, this model also assumes that the input indices are conditionally independent once again making a brittle assumption.

Despite these constraints, the study offered a structured path toward a more holistic

multi-indicator integration and contributed to the validity of using DNBCs for composite drought indicators [34].

Review of In-Situ and Remote Sensing-Based Indices and Their Applicability for Integrated Drought Monitoring in South Africa

This study aimed to critically assess the performance and applicability of both in-situ and remote sensing-based drought indices for integrated drought monitoring in South Africa. Its objectives were to evaluate eight widely used indices and to determine which are most suitable for South Africa's highly variable climate. These eight indices were: PDSI, SWSI, VCI, SPI, SPEI, SSI, SGI, and GRACE-based indices. A further aim was to test the hypothesis that no single index can adequately capture all aspects of meteorological, agricultural, and hydrological drought.

They followed the World Meteorological Organisation's (WMO) 2016 guidelines for drought indicator assessment. They used five evaluation criteria focusing on capability, sensitivity, data requirements, computational simplicity, and versatility for integration. The review drew from published studies in South Africa and other regions with similar climate characteristics. The indices were chosen based on surveys, while their feasibility was assessed against the evaluation framework mentioned.

The findings demonstrated that the PDSI and SWSI are not feasible to obtain in South Africa due to their high complexity with regards to data requirements. However, SPI, SPEI, VCI, SSI, and SGI were identified as the most feasible candidates for integrated drought monitoring because of their simplicity and adaptability. Regardless, calculation issues remain, for example, there is no consensus on the most suitable probability distribution functions (PDF) for the calculations of SSI and SGI, with the most commonly used Gamma distribution performing poorly in South African catchments. Some alternative distributions showed improved results but inconsistencies persisted. Finally, the review recommended exploring multivariate approaches that combine SPI, SPEI, VCI, SSI, and SGI, while also noting the potential of GRACE-based indices, particularly with regards to groundwater, in order to compensate for the country's limited groundwater records.

The study transparently noted some important limitations. Data availability constraints undermine the feasibility of effective indices such as the PDSI and SGI, while the scarcity, or absence, of groundwater records limits applications. PDF selection for SSI and SGI remains uncertain given the climate variation in South Africa. The authors identified key research gaps within the nation, including the need for multivariate index testing and more exploration of GRACE-based products.

Ultimately, the review emphasised that integrated approaches, underpinned by sensitivity analysis and comparative testing, are required to strengthen drought monitoring in South Africa's complex climatic landscape [32].

Developing a Composite Drought Indicator Using PCA Integration of CHIRPS Rainfall, Temperature, and Vegetation Health Products for Agricultural Drought Monitoring in New Mexico

The objective of this study was to construct a Composite Drought Indicator for New Mexico, the so called CDI-NM, by integrating multiple variables through Principal Component Analysis (PCA). The research sought to provide a drought monitoring tool capable of identifying historical drought events, while also quantifying drought extent across the state. The study combined satellite-derived rainfall, temperature, and vegetation health products to demonstrate the effectiveness of PCA and to investigate drought impacts on agricultural production.

The methodology focused on New Mexico which is an agriculturally important US state and is vulnerable to varying climates. Four input datasets spanning 2003–2019 were incorporated: CHIRPS rainfall data, MODIS Land Surface Temperature (LST), Smoothed Normalized Difference Vegetation Index (SMN), and Vegetation Condition Index (VCI). PCA was conducted independently for each month, with suitability being validated using Kaiser-Meyer-Olkin and Bartlett's tests. They tested their model output by comparing it against SPI-3 and by correlating it with the annual variations in the yields of wheat, corn, peanuts, and cotton.

The results indicated that CDI-NM showed strong agreement with SPI-3, effectively capturing major drought events in 2003, 2011–2013, and 2018. Additionally, the showed their CDI-NM had strongly negative correlations with yields for corn (-0.68) and wheat (-0.63), while having a weaker correlation with cotton (-0.20). This reflects greater drought tolerance for cotton. Relationships between input variables were also consistent with expectation, as positive correlation was seen between VCI and rainfall (0.78) and negative correlation with LST (-0.43). Finally, the indicator demonstrated more natural variations than SPI, suggesting improved ability at capturing agricultural drought.

Despite these achievements, several limitations were identified. The method of PCA relies on linear assumptions, temporal stationarity, and is sensitivity to scaling. The 17-year dataset is inherently limited with regards to long-term generalisability. Some data-related uncertainties further constrained precision. Redundancy between NDVI-derived SMN and VCI also posed risks of over-representation of vegetation conditions. Moreover, the study did not conduct sensitivity testing of PCA-derived weights, leaving gaps in applicability. The authors highlighted the need for longer datasets, uncertainty assessments, and more advanced dimensionality reduction techniques to strengthen the reliability of composite indicators for drought monitoring [35].

Conclusion

The literature shows that probabilistic approaches are promising for capturing drought's complex nature, offering an advantage over traditional indices. Although South Korean research offers a strong blueprint, the scarcity of composite indicator development in South Africa reveals a significant research gap. This project seeks to bridge this gap by tailoring a DNBC to South Africa.

Chapter 3

Methods

3.1. Data Acquisition

The development of a composite drought indicator requires careful selection of input variables that capture the different aspects of drought. Three indices were selected: the Standardised Precipitation Index (SPI) to represent meteorological drought, the Streamflow Drought Index (SDI) to represent hydrological drought, and the Normalised Difference Vegetation Index (NDVI) as a proxy for agricultural drought. These indices were chosen based on their widespread use in literature and the availability of data. Data scarcity is a challenge in South Africa, as openly accessible, long and consistent drought-related datasets are limited. Consequently, the choice of indices attempts to strike a balance between theory and pragmatic constraints **??**.

??

3.1.1. Sources

To compute the SPI, monthly rainfall data was obtained from the University of Cape Town (UCT) dataset, which covers the period 1979–2019 (Dataset **??**). The dataset provides rainfall values at the station level, offering a high degree of spatial granularity across South Africa.

For the SDI, daily streamflow records were obtained from the Department of Water and Sanitation (DWS), which maintains audited historic data regarding hydrology (Dataset **??**). These daily records were averaged to monthly to calculate the target index.

To obtain the NDVI, the NOAA Climate Data Record (CDR) of AVHRR Normalised Difference Vegetation Index (NDVI), Version 5 was used (Dataset **??**). The dataset spans the period 1981–2025 and is provided in global NetCDF format. For the purposes of this study, only the South African subset was extracted. This required targeted downloading and filtering, given the large size of the global dataset.

3.1.2. Preprocessing

To prepare the indices for model input, several preprocessing steps were performed:

1. **Time Period Alignment:** All datasets were resampled or aggregated to a common monthly resolution.

2. **Area Alignment:** For station-based datasets, like rainfall and streamflow, records were harmonised by selecting stations with consistent temporal coverage. For NDVI, gridded data was averaged over the area of choice.

3. **Brief Exploration Of Data:** The data sets were analysed to identify and issues in the data such as missing/null values, format consistency, validity, etc. No problems were found

3.2. Index Calculation

The selected datasets were subsequently transformed into drought indices using established methodologies. The SPI was derived from rainfall anomalies through standardisation against a long-term climatology. The SDI was computed by standardising streamflow anomalies relative to long-term hydrological records. The NDVI, while not a drought index in its raw form, was used to reflect vegetation stress associated with agricultural drought conditions. References to the seminal works underlying these methodologies are provided in the bibliography.

3.2.1. SPI

The SPI is based on the statistical normalisation of accumulated precipitation over a specified time window. Precipitation values are first fitted to a probability distribution, commonly the gamma distribution, and then transformed into a standard normal distribution. This yields an index with mean zero and unit variance, allowing for direct interpretation of drought severity across different temporal scales.

3.2.2. SDI

The SDI extends the concept of the SPI to streamflow. It is calculated by aggregating streamflow over a specified time window and standardising it against long-term flow records. Positive values of SDI indicate above-normal hydrological conditions, while negative values reflect hydrological drought. The SDI is particularly relevant in South Africa, where surface water storage and river systems play a critical role in drought impact and management.

3.2.3. NDVI

The NDVI is derived from remotely sensed reflectance in the red and near-infrared bands of the electromagnetic spectrum. Vegetated surfaces typically absorb red light for

photosynthesis and reflect near-infrared light, making the NDVI an effective indicator of vegetation health. While NDVI is not a drought index per se, reductions in NDVI are widely used to monitor agricultural drought stress, as vegetation is sensitive to deficits in soil moisture and precipitation.

3.2.4. Discretisation of Indices

The SPI, SDI, and NDVI are all continuous variables. However, the Dynamic Naïve Bayes Classifier requires discrete inputs. Accordingly, each index was discretised into categorical bins based on thresholds reported in the literature and common practice. For example, SPI values are often classified into categories such as "extremely dry," "moderately dry," and "normal." This discretisation not only facilitates model implementation but also aligns with the interpretive categories commonly employed in drought monitoring. A more detailed account of the discretisation procedure is provided in Section **??**.

Table 3.1: Discretisation thresholds for drought indices.

Category	SPI / SDI	NDVI Anomaly	Interpretation
Severe Drought	≤ -1.5	≤ -1.5	Extreme vegetation/hydrological stress
Moderate Drought	$-1.5 < x \leq -0.5$	$-1.5 < x \leq -0.5$	Sustained but moderate deficit
Normal	$-0.5 < x < 0.5$	$-0.5 < x < 0.5$	Near-average conditions
Moderate Wet	$0.5 \leq x < 1.5$	$0.5 \leq x < 1.5$	Above-average moisture/greenness
Severe Wet	≥ 1.5	≥ 1.5	Flooding risk or excessive rainfall

3.3. Model Development

3.3.1. Model Design

Defining the Random Variables

The proposed DNBC is constructed in a general form with N input variables observed across T discrete time steps. All random variables (RVs) in the model are treated as discrete.

The first set of RVs corresponds to the latent drought states at each time step, denoted by

$$S_t \in \{1, 2, \ldots, m\}, \quad t = 1, \ldots, T,$$

where m represents the number of possible drought states. This value of m is not fixed, but will rather be determined via model selection.

The second set of RVs corresponds to the observed input variables, denoted by

$$A_t^{(n)} \in \{1, 2, \ldots, C_n\}, \quad n = 1, \ldots, N, \quad t = 1, \ldots, T,$$

where C_n is the cardinality of the n-th input variable. In this project, these inputs are the indices used to represent different aspects of drought:

$$\text{SPI} \equiv A_t^{(1)}, \quad \text{SDI} \equiv A_t^{(2)}, \quad \text{NDVI} \equiv A_t^{(3)}.$$

These observed indices constitute the data set $\mathcal{D}$.

For clarity, we define the following notation which will be used throughout the model formulation:

$$\mathbf{S}_{1:T} = \{S_1, S_2, \ldots, S_T\}, \quad A_{1:T} = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_T\},$$

where each $\mathbf{A}_t = \{A_t^{(1)}, A_t^{(2)}, \ldots, A_t^{(N)}\}$.

The total number of latent state nodes is therefore T, while the number of observed input nodes is $T \times N$. A summary of the random variables, their number of nodes, and their cardinality is provided in Table 3.2.

Table 3.2: Summary of random variables in the model

Name	Number of Nodes	Cardinality
Latent drought state S_t	T	m
General input variable $A_t^{(n)}$	$T \times N$	C_n

Graphical Structure & Assumptions

Figure 3.1 below displays the model diagram for a DNBC for T time steps and N input variables.

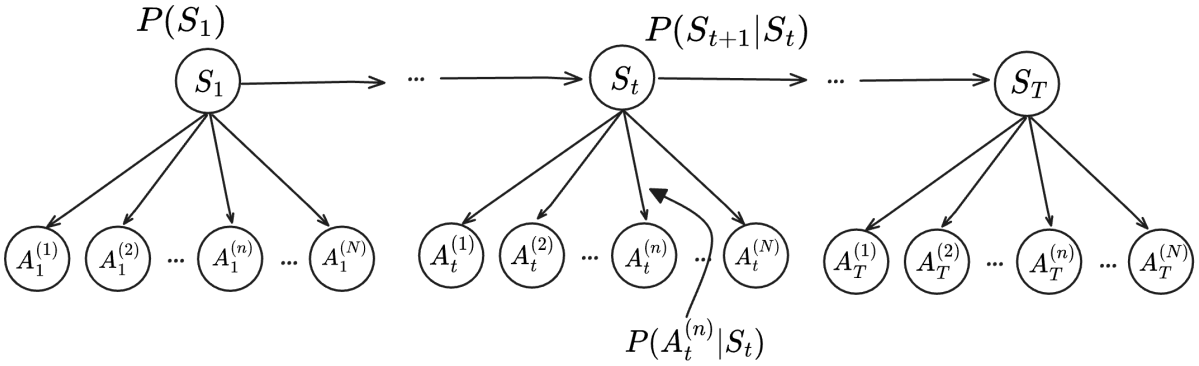

Figure 3.1: The Dynamic Naïve Bayes Classifier (DNBC) can be represented as a Bayesian network unfolding over time. At each time step t, a latent drought state S_t is modelled as a discrete random variable that governs the latent structure, while the observed input variables $\mathbf{A}_t = \{A_t^{(1)}, A_t^{(2)}, \ldots, A_t^{(N)}\}$ are each solely dependent on S_t

It is important to note the inherent limitations of this model, that being:

(i) The dynamic process of the state sequence S_t follows a first-order Markov chain. This means the state at time $t+1$ is conditionally dependent only on the state at time t.

(ii) The dynamic process is stationary, implying that the transition probabilities between states are constant over time.

(iii) For each time step t, the model assumes conditional independence among the input variables $\mathbf{A}_t$ given the corresponding hidden drought state S_t.

Joint Distribution

The joint probability distribution of the observed variables and latent states in the DNBC can be expressed as:

$$\begin{aligned} p(S_1,&S_2,\ldots,S_T,A_1^{(1)},A_1^{(2)},\ldots,A_1^{(N)},A_2^{(1)},\ldots,A_T^{(N)}) \\ &= p(S_1,S_2,\ldots,S_T,\mathbf{A}_1,\mathbf{A}_2,\ldots,\mathbf{A}_T) \\ &= p(\mathbf{S}_{1:T},A_{1:T}) \\ &= p(S_1)\cdot\prod_{t=1}^{T-1}p(S_{t+1}\mid S_t)\cdot\prod_{n=1}^{N}\prod_{t=1}^{T}p(A_t^{(n)}\mid S_t) \end{aligned} \tag{3.1}$$

The following factorisation is possible due to Assumption (iii) of the DNBC and will become useful at a later stage.

$$\begin{aligned} p(\mathbf{A}_t\mid S_t) &= p(A_t^{(1)},A_t^{(2)},\ldots,A_t^{(N)}\mid S_t) \\ &= p(A_t^{(1)}\mid S_t)p(A_t^{(2)}\mid S_t)(A_t^{(N)}\mid S_t) \\ &= \prod_{n=1}^{N}p(A_t^{(n)}\mid S_t) \end{aligned} \tag{3.2}$$

Parameterising the Model

The DNBC is fully specified by three sets of parameters, that being the prior, transition and emission probabilities.

Prior Probabilities: The initial distribution over the latent drought S_1.
The factor table for the priors is show below in Table 3.3:

Table 3.3: Priors Factor Table

S_1	$p(S_1)$
1	π_1
2	π_2
$\vdots$	$\vdots$
m	π_m

where π_i is the probability that the system begins in state i.

$$\pi_i \equiv p(S_1 = i),$$

Transition Probabilities: defines the likelihood of moving to a new hidden state given the current hidden state.
The factor table as well as the transition matrix P^1 is shows below in Table 3.4:

Table 3.4: Transition Factor Table & Transition Matrix

S_t	S_{t+1}	$p(S_{t+1} \mid S_t)$
1	1	$a_{1,1}$
1	2	$a_{1,2}$
⋮	⋮	⋮
1	m	$a_{1,m}$
2	1	$a_{2,1}$
2	2	$a_{2,2}$
⋮	⋮	⋮
m	m	$a_{m,m}$

$$\equiv \quad P^1 = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,m} \\ a_{2,1} & a_{2,2} & \dots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \dots & a_{m,m} \end{bmatrix}$$

Here, $a_{i,j}$ represents the probability of transitioning from state i at time t to state j at time $t+1$.

$$a_{i,j} \equiv p(S_{t+1} = j \mid S_t = i).$$

Note as well that transition matrix's rows sum to 1, ie. $\sum_{j=1}^{m} a_{i,j} = 1$ for all i.

Emission Probabilities: Defines the likelihood of observing a particular inout variable, given that the system is in a specific hidden state.
Once again, the factor table for the emission probabilities is show below in Table 3.5

Table 3.5: Emission Factor Table

$A_t^{(n)}$	S_t	$p(A_t^{(n)} \mid S_t)$
1	1	$b_1^{(n)}(1)$
1	2	$b_2^{(n)}(1)$
⋮	⋮	⋮
1	m	$b_m^{(n)}(1)$
2	1	$b_1^{(n)}(2)$
2	2	$b_2^{(n)}(2)$
⋮	⋮	⋮
C_n	m	$b_m^{(n)}(C_n)$

Where, $b_i^{(n)}(j)$ is the likelihood of observing input variable n take on the value j, given

the its corresponding hidden drought state is equal to i

$$b_i^{(n)}(j) \equiv p(A_t^{(n)} = j \mid S_t = i).$$

These parameters encode how the drought indicators behave under each latent drought state. Taken together, the parameter set fully determines the DNBC. It is important to note that due to the parameters being time independent, as the model assumes stationarity, the rules governing drought state transitions and emissions are invariant across time.

3.3.2. Inference

In this section, inference for the DNBC is developed under the assumption that the parameters Θ are known and the input variables $A_{1:T}$ are observed. Since the attributes are not random at this stage, the task becomes trying to infer the distribution of the hidden drought states:

$$p(\mathbf{S}_{1:T} \mid A_{1:T}, \Theta),$$

This will later be used for the E-step in the EM algorithm.

The inference procedure is carried out using the Junction Tree (JT) framework, which provides exact inference. Messages are propagated through the tree, beginning at the leaf clusters and moving inward [36].
Figure 3.2 illustrates the JT structure associated with the DNBC.

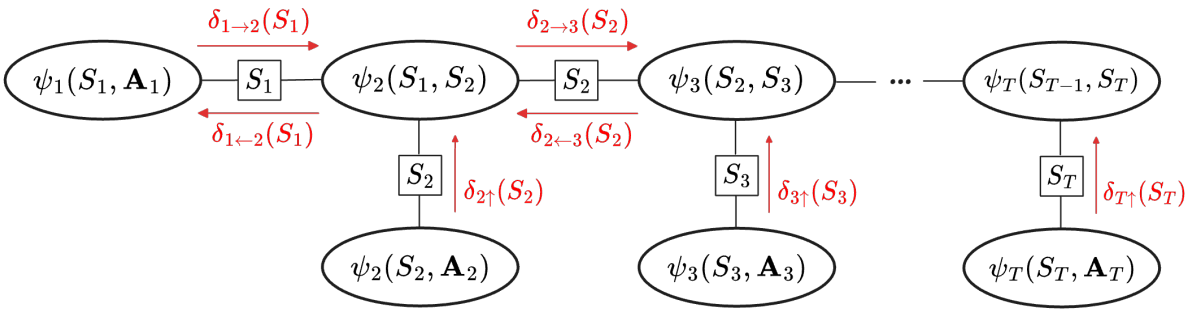

Figure 3.2: Junction Tree representation of the DNBC. Each cluster groups together latent state variables and observed attributes, with sepsets defined along the edges. Messages are propagated through the tree to perform exact inference.

It is useful to note the factorisation of the observed attribute RVs at Equation 3.2. The cluster potentials are summarised in Table 3.6.

Message Passing

Messages are defined between clusters, with sepsets given by the product of messages between clusters.

Table 3.6: Cluster potentials for the DNBC. Each potential corresponds either to a state transition or to a state-attribute relationship.

–	$\psi_1(S_1, \mathbf{A}_1) = p(S_1)p(\mathbf{A}_1 \mid S_1)$
$\psi_2(S_1, S_2) = p(S_2 \mid S_1)$	$\psi_2(S_2, \mathbf{A}_2) = p(\mathbf{A}_2 \mid S_2)$
$\vdots$	$\vdots$
$\psi_t(S_{t-1}, S_t) = p(S_t \mid S_{t-1})$	$\psi_t(S_t, \mathbf{A}_t) = p(\mathbf{A}_t \mid S_t)$
$\vdots$	$\vdots$
$\psi_T(S_{T-1}, S_T) = p(S_T \mid S_{T-1})$	$\psi_T(S_T, \mathbf{A}_T) = p(\mathbf{A}_T \mid S_T)$

Upward messages: Because the attributes are observed, upward messages collapse to the corresponding likelihood terms.

$$\begin{aligned}\delta_{t\uparrow}(S_t) &= \sum_{\mathbf{A}_t} \psi_t(S_t, \mathbf{A}_t) \\ &= \sum_{\mathbf{A}_t} p(\mathbf{A}_t \mid S_t) \\ &= p(\mathbf{A}_t \mid S_t)\end{aligned}$$

since marginalisation over the observed attributes reduces to their likelihood.

Rightward messages: Rightward propagation starts at the leftmost cluster and moves forward in time:

$$\begin{aligned}\delta_{1\to 2}(S_1) &= \sum_{\mathbf{A}_1} \psi_1(S_1, \mathbf{A}_1) \\ &= \sum_{\mathbf{A}_1} p(S_1)p(\mathbf{A}_1 \mid S_1) \\ &= p(S_1)p(\mathbf{A}_1 \mid S_1) \qquad (3.3)\end{aligned}$$

$$\begin{aligned}\delta_{t\to t+1}(S_t) &= \sum_{S_{t-1}} \psi_t(S_{t-1}, S_t)\delta_{t-1\to t}(S_{t-1})\delta_{t\uparrow}(S_t) \\ &= \sum_{S_{t-1}} p(S_t \mid S_{t-1})\delta_{t-1\to t}(S_{t-1})p(\mathbf{A}_t \mid S_t) \\ &= p(\mathbf{A}_t \mid S_t) \sum_{S_{t-1}} p(S_t \mid S_{t-1})\delta_{t-1\to t}(S_{t-1}) \qquad (3.4)\end{aligned}$$

Leftward messages. Similarly, leftward propagation begins at the final cluster and proceeds backward:

$$\delta_{T-1\leftarrow T}(S_{T-1}) = \sum_{S_T} p(S_T \mid S_{T-1}) p(\mathbf{A}_T \mid S_T), \tag{3.5}$$

$$\delta_{t-1\leftarrow t}(S_{t-1}) = \sum_{S_t} p(S_t \mid S_{t-1}) \delta_{t\leftarrow t+1}(S_t) p(\mathbf{A}_t \mid S_t). \tag{3.6}$$

Remarks

In this framework, the clusters of primary interest are $\psi_t(S_t, S_{t+1})$ and the sepsets $\mu_{t,t+1}(S_t)$, which directly contribute to the computation of $p(\mathbf{S}_{1:T} \mid A_{1:T}, \Theta)$. As a result, downward messages (e.g., from $\psi_t(S_{t-1}, S_t)$ to $\psi_t(S_t, \mathbf{A}_t)$) are not of interest.
Finally, it is worth noting that for JTs, since the underlying graph is a tree, message passing is exact. We follow a specific message-passing ordering of the standard Belief Propagation algorithm, which is guaranteed to converge to the exact marginals.

Forward–Backward Algorithm

At this point, it is natural to highlight the connection between the JT approach described above and the more classical algorithms for HMMs along with their variants. Readers familiar with the literature will recognise that the message passing operations we performed are precisely the equivalent to the well-known *forward–backward equations* [37–39].
The forward and backward recursions applied to the proposed model are shown below:

Forward:

$$\begin{aligned}
&\text{Define:} \quad \alpha_t^k = p(A_{1:t}, S_t = i) \\
&\text{Init:} \quad \alpha_1^k = p(S_1 = k) p(\mathbf{A}_1 \mid S_1 = k) \\
&\text{Iteration:} \quad \alpha_t^k = p(\mathbf{A}_t \mid S_t = k) \sum_{i=1}^{m} \alpha_{t-1}^i \cdot p(S_t = k \mid S_{t-1} = i)
\end{aligned} \tag{3.7}$$

Backward:

$$\begin{aligned}
&\text{Define:} \quad \beta_t^k = p(A_{1:t}, S_t = i) \\
&\text{Init:} \quad \beta_T^k = 1 \quad \forall\, k \\
&\text{Iteration:} \quad \beta_t^k = \sum_{i=1}^{m} p(S_{t+1} = i \mid S_t = k) \cdot p(\mathbf{A}_{t+1} \mid S_{t+1} = i) \cdot \beta_{t+1}^i
\end{aligned} \tag{3.8}$$

Remarks on Forward–Backward and Baum–Welch

The messages passed in the JT (Equations 3.3 - 3.6) coincide with the α and β recursions in Equations 3.7–3.8. The distinction is thus in presentation alone. The JT framework is a generalisation for arbitrary graphical models, whereas the forward–backward is the special case formulation for the structure of HMMs [40].

It is worth emphasising the parallel between the JT messages and the forward–backward quantities. The forward recursion $\alpha_t^k = p(A_{1:t}, S_t = k)$ and the backward recursion $\beta_t^k = p(A_{t+1:T} \mid S_t = k)$ are algebraically equivalent to the inward and outward sum–product messages in the JT 3.2. When inward and outward messages are combined at a cluster or sepset, the resulting posterior marginals $p(S_t \mid A_{1:T})$ and pairwise marginals $p(S_t, S_{t+1} \mid A_{1:T})$ coincide with the responsibilities computed from Baum-Welch. Thus, the JT message-passing procedure and the forward–backward algorithm produce identical posterior marginals. These results will be of interest in the following sub section for parameter estimation [**?**, 34, 39, 40].

In summary, the JT formulation highlights the structural perspective, while forward–backward and Baum–Welch remain the traditional algorithms in the literature. Both views are mathematically equivalent and lead to the same computations.

3.3.3. Parameter Estimation

Parameter estimation for the DNBC is carried out using the Expectation–Maximization (EM) algorithm [41]. We distinguish between the hidden variables, observed data, and model parameters as follows:

$$\mathcal{H} = (S_t)_{t=1}^T$$
$$\mathcal{D} = (\mathbf{A}_t)_{t=1}^T$$
$$\Theta = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$$

Where,

$$\boldsymbol{\theta}_1 = \{\pi_1, \pi_2, \ldots, \pi_m\} \equiv \text{Priors Probabilities}$$
$$\boldsymbol{\theta}_2 = \{a_{i,j} \mid i, j = 1, \ldots, m\} \equiv \text{Transition Probabilities}$$
$$\boldsymbol{\theta}_3 = \left\{ b_i^{(n)}(j) \,\middle|\, i = 1, \ldots, m;\ n = 1, \ldots, N;\ j = 1, \ldots, C_n \right\} \equiv \text{Emission Probabilities}$$

The EM algorithm iteratively alternates between two steps:

1. E-Step

In this step, we hold Θ fixed and compute the posterior distribution over the hidden states:

$$\begin{aligned} q(\mathcal{H}) &= p(\mathcal{H} \mid \mathcal{D}, \Theta) \\ &= p(\mathbf{S}_{1:T} \mid A_{1:T}, \Theta) \end{aligned} \tag{3.9}$$

This corresponds directly to the inference problem, as previously discussed in Section 3.3.2.

2. M-Step

Next, with q fixed, we maximise the variational lower bound

$$\mathscr{L}(q, \Theta) = \sum_{\mathcal{H}} q(\mathcal{H}) \cdot \log \left(\frac{p(\mathcal{D},\ \mathcal{H} \mid \Theta)}{q(\mathcal{H})} \right)$$

with respect to Θ.

Equivalently, this requires solving

$$\begin{aligned} \Theta &= \underset{\Theta}{\operatorname{argmax}}\ \mathcal{Q}(\Theta) \\ &= \underset{\Theta}{\operatorname{argmax}} \sum_{\mathcal{H}} q(\mathcal{H}) \cdot \log\ p(\mathcal{D}, \mathcal{H} \mid \Theta) \end{aligned} \tag{3.10}$$

The inner term, $\log\ p(\mathcal{D}, \mathcal{H} \mid \Theta)$, is simply the log of the joint distribution introduced in Equation 3.1. Expanding this expression yields:

$$\begin{aligned} p(A_{1:T}\, \mathbf{S}_{1:T} \mid \Theta) = \log\ & p(S_1 \mid \boldsymbol{\theta}_1) \\ & + \sum_{t=1}^{T-1} \log\ p(S_{t+1} \mid S_t,\ \boldsymbol{\theta}_2) \\ & + \sum_{n=1}^{N} \sum_{t=1}^{T} \log\ p(A_t^{(n)} \mid S_t,\ \boldsymbol{\theta}_3) \end{aligned}$$

Substituting this into $\mathcal{Q}(\Theta)$ and carefully reorganising terms allows us to isolate contributions from priors, transitions, and emissions. Since all RVs are discrete, probabilities translate directly into parameterised forms, and the optimisation decouples naturally across $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, and $\boldsymbol{\theta}_3$.

$$\begin{aligned}\mathcal{Q} &= \sum_{\mathcal{H}} q(\mathcal{H}) \cdot \log\, p(\mathcal{D}, \mathcal{H} \mid \Theta) \\ &= \sum_{\mathcal{H}} q(\mathcal{H}) \cdot \Big[\log\, p(S_1 \mid \boldsymbol{\theta}_1) \\ &\qquad + \sum_{t=1}^{T-1} \log\, p(S_{t+1} \mid S_t,\, \boldsymbol{\theta}_2) \\ &\qquad + \sum_{n=1}^{N} \sum_{t=1}^{T} \log\, p(A_t^{(n)} \mid S_t,\, \boldsymbol{\theta}_3) \Big]\end{aligned}$$

We then multiply $\sum_{\mathcal{H}} q(\mathcal{H})$ through, understanding that $\mathcal{H} = (S_1, \ldots, S_T)$

$$\begin{aligned}&= \sum_{S_1,\ldots,S_T} \log\, p(S_1 \mid \boldsymbol{\theta}_1)\, q(S_1, \ldots, S_T) \\ &\quad + \sum_{S_1,\ldots,S_T} \sum_{t=1}^{T-1} \log\, p(S_{t+1} \mid S_t,\, \boldsymbol{\theta}_2)\, q(S_1, \ldots, S_T) \\ &\quad + \sum_{S_1,\ldots,S_T} \sum_{n=1}^{N} \sum_{t=1}^{T} \log\, p(A_t^{(n)} \mid S_t,\, \boldsymbol{\theta}_3)\, q(S_1, \ldots, S_T)\end{aligned}$$

$$\begin{aligned}&= \sum_{S_1} \log\, p(S_1 \mid \boldsymbol{\theta}_1)\, q(S_1) + \sum_{S_2,\ldots,S_T} q(S_2, \ldots, S_T) \\ &\quad + \sum_{t=1}^{T-1} \sum_{S_t, S_{t+1}} \log\, p(S_{t+1} \mid S_t,\, \boldsymbol{\theta}_2)\, q(S_t, S_{t+1}) + \sum_{\substack{S_1,\ldots,S_T \\ \setminus S_t,\, S_{t+1}}} q(S_1, \ldots, S_T) \\ &\quad + \sum_{n=1}^{N} \sum_{t=1}^{T} \sum_{S_t} \log\, p(A_t^{(n)} \mid S_t,\, \boldsymbol{\theta}_3)\, q(S_t) + \sum_{\substack{S_1,\ldots,S_T \\ \setminus S_t}} q(S_1, \ldots, S_T)\end{aligned}$$

Since the goal is to optimise w.r.t $\Theta = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$, all the terms not involving Θ can be dropped, whilst also using the result $\sum_{S_t} p(S_t) = \sum_{i=1}^{m} p(S_t = i)$

$$\begin{aligned}&= \sum_{i=1}^{m} \log\, p(S_1 = i \mid \boldsymbol{\theta}_1)\, q(S_1 = i) \\ &\quad + \sum_{t=1}^{T-1} \sum_{i=1}^{m} \sum_{j=1}^{m} \log\, p(S_{t+1} = j \mid S_t = i,\, \boldsymbol{\theta}_2)\, q(S_t = i, S_{t+1} = j) \\ &\quad + \sum_{n=1}^{N} \sum_{t=1}^{T} \sum_{i=1}^{m} \log\, p(A_t^{(n)} \mid S_t = i,\, \boldsymbol{\theta}_3)\, q(S_t = i)\end{aligned}$$

This representation can be expressed in terms of the model parameters. Because all random variables are discrete, the probabilities naturally reduce to combinations of these parameters.

$$
\begin{aligned}
&= \sum_{i=1}^{m} q(S_1 = i) \log \pi_i \\
&\quad + \sum_{t=1}^{T-1} \sum_{i=1}^{m} \sum_{j=1}^{m} q(S_t = i, S_{t+1} = j) \log a_{i,j} \\
&\quad + \sum_{t=1}^{T} \sum_{i=1}^{m} q(S_t = i) \sum_{n=1}^{N} \log b_i^{(n)}(A_t^{(n)})
\end{aligned}
$$

Each of the target parameters are now separated into their own terms and thus can be easily optimised in isolation. This yields the standard re-estimation updates [42, 43]:

$$\boxed{\pi_i^{\text{new}} = q(S_1 = i)} \tag{3.11}$$

$$\boxed{a_{i,j}^{\text{new}} = \frac{\sum_{t=1}^{T-1} q(S_t = i, S_{t+1} = j)}{\sum_{t=1}^{T-1} q(S_t = i)}} \tag{3.12}$$

$$\boxed{b_i^{(n)}(j)^{\text{new}} = \frac{\sum_{t=1}^{T} q(S_t = i) \cdot \mathbf{1}(A_t^{(n)} = j)}{\sum_{t=1}^{T} q(S_t = i)}} \tag{3.13}$$

3.3.4. Model Selection

Model selection will involve determining the appropriate cardinality of each latent drought states S_t, that is, determining the value of m. When selecting m, a balance must be struck between model complexity and goodness of fit, as a larger value of m gives the model a greater ability to capture subtle drought dynamics but risks overfitting. On the other hand, a smaller number may be too restrictive to reflect the underlying processes.

To guide this choice, three complementary criteria are applied: the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the maximised log-likelihood of the fitted model. These are given by

$$AIC = -2 \cdot \log L(\Theta) + 2p, \tag{3.14}$$

$$BIC = -2 \cdot \log L(\Theta) + p \cdot \log k, \tag{3.15}$$

where $L(\Theta)$ is the maximised value of the likelihood function, p is the number of free

parameters in the model, and k is the number of data points.

The philosophy underlying these criteria is rooted in Occam's razor, which is often phrased as "the simplest explanation is usually the best one". AIC and BIC both balance model fit against complexity, but with differing severity. BIC applies a stronger penalty on complexity and is thus generally considered more consistent with Occam's razor [44]. Thus, the framework for selecting m as follows:

1. **Primary:** select the model with the lowest BIC, penalising unnecessary complexity.
2. **Secondary:** use AIC to cross-check results.
3. **Tertiary:** inspect the log-likelihood curve. If $\log L(\Theta)$ improves only marginally as m increases, the simpler model is preferred (the so-called "elbow rule").

In practice, model selection is performed by sweeping across candidate values of m, fitting a model for each case, and comparing their AIC, BIC, and log-likelihood values. The final choice of m seeks to minimise both AIC and BIC while ensuring that the likelihood $L(\Theta)$ does not deteriorate substantially.

Choice of k

The term k in (3.15) represents the number of data points. Following common practice in the literature and implementation libraries such as the `seqHMM` package in R [45], k is calculated as:

$$k = T \times N,$$

Number of Free Parameters p

The number of free parameters p corresponds to the model's degrees of freedom. This includes contributions from the prior probabilities 3.3, the transition probabilities 3.4, and the emission probabilities 3.5. It is widely accepted in the literature and implementations regarding HMMs and its variants [45, 46] that

$$\begin{aligned} p &= (m-1) + m(m-1) + \sum_{n=1}^{N} m(C_n - 1) \\ &= m^2 - 1 + m \sum_{n=1}^{N} (C_n - 1), \end{aligned}$$

Log-Likelihood Estimation

The third component of model selection is the log-likelihood, $\ell(\Theta)$, which measures the probability of the observed data under the model parameters:

$$\ell(\Theta) = p(A_{1:T}^{\text{obs}} \mid \Theta).$$

This likelihood can be evaluated efficiently using the forward algorithm. Recall that the forward variable is defined as

$$\alpha_t^k = p(S_t = k, A_{1:t} \mid \Theta),$$

The overall likelihood is then simply obtained by marginalising over the latent state at the final time step:

$$\begin{aligned}\sum_{i=1}^{m} \alpha_T^i &= \sum_{S_T} p(A_{1:T}, S_T \mid \Theta) \\ &= p(A_{1:T} \mid \Theta) \;=\; \ell(\Theta).\end{aligned}$$

Although the derivation via the forward algorithm is given for clarity, it has been established that it is equivalent to the rightward message-passing procedure in the JT approach (Section 3.3.2). For this approach, an additional downward message $\delta_{\downarrow T}(S_T)$ must be computed at the final cluster to obtain the posterior $\psi_T(S_T, \mathbf{A}_T)$. Marginalising out S_T from this posterior yields the desired likelihood. It is useful to see Figure 3.2.

Create a final draft for this:

3.4. Model Implementation

3.4.1. Programming Environment & Tools

All aspects of the DNBC model were implemented in `C++`, primarily chosen for its computational efficiency and the availability of the `emdw` library. This library provides robust functionality for probabilistic graphical models. The `C++` implementation handled the construction of factors, junction tree message passing, parameter estimation, model selection, and extraction of posterior outputs.

Python was used to complement this workflow, particularly for data-related tasks such as raw data extraction, preprocessing into model inputs, postprocessing of model outputs, and visualisation of results. This division allowed `C++` to focus on core model computation while Python streamlined data management and analysis.

3.4.2. Data Pipeline Implementation

The data pipeline was designed to translate raw climate data into discretised indices that serve as inputs to the model. The visualisation of this pipeline and its flow is shown in Figure 3.3 At a high level, the process consisted of:

1. **Data Collection:** Acquiring raw climate and vegetation data.

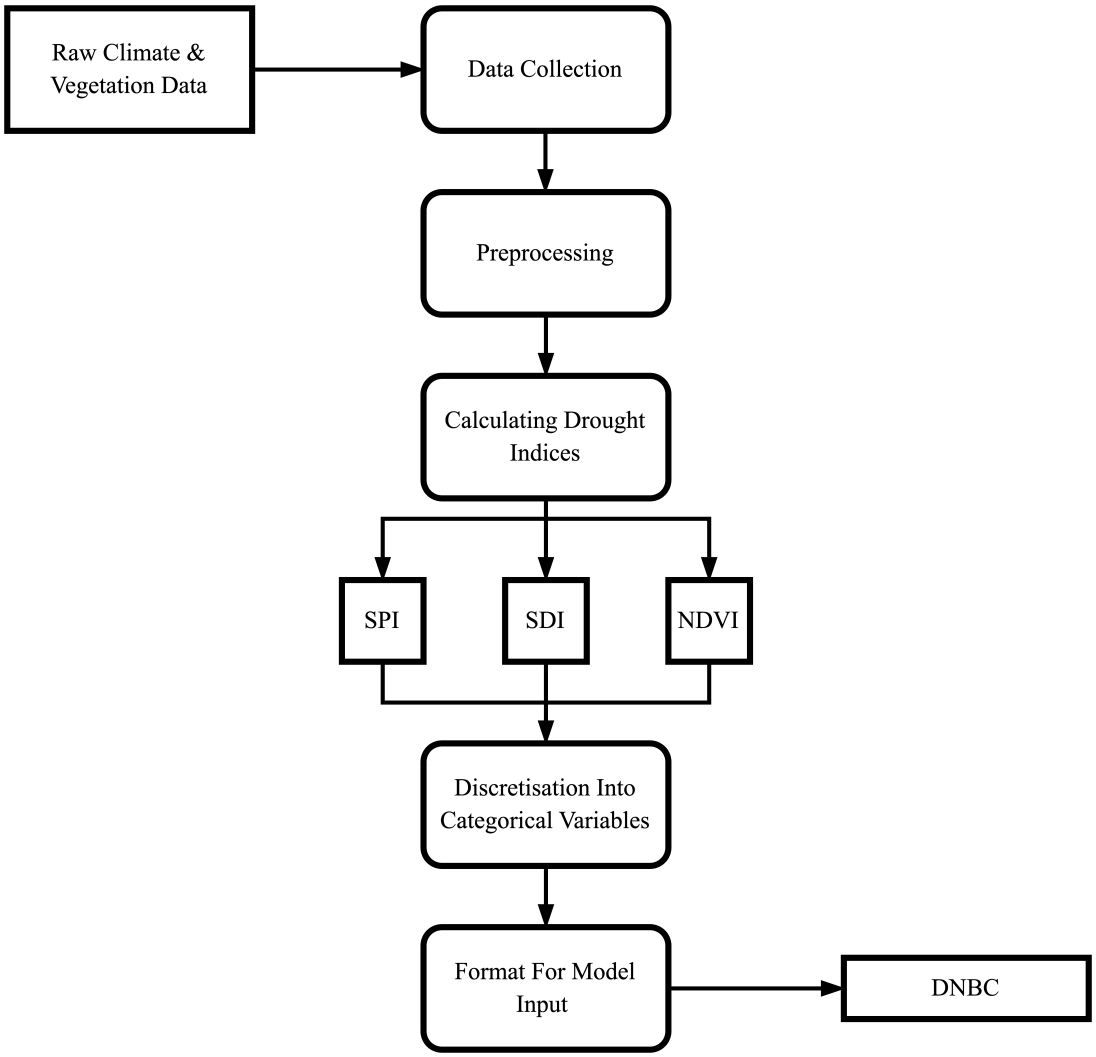

Figure 3.3: Data pipeline from raw climate and vegetation data to discretised indices used as DNBC inputs. The pipeline consists of five stages: data collection, preprocessing, index calculation, discretisation, and input formatting.

2. **Preprocessing:** Handle missing values, time and space alignment, ensuring data consistency, etc.

3. **Index Calculation:** Input indices (SPI, SDI, and NDVI) were calculated following formulas given in Section 3.2.

4. **Discretisation:** Convert continuous indices into categorical for DNBC input.

5. **Input Formatting:** Finally, these discretised indices are formatted and saved as a CSV file, ready for model ingestion with `C++`.

This pipeline was implemented using Python.

3.4.3. Model Implementation

Implementation of the DNBC was achieved through two main functions: `runEM` and `modelSelection`.

runEM

The `runEM` function performed parameter estimation using the EM algorithm, paired with exact inference offered by the JT methodology:

1. Random initialisation of parameters (sampled from a Gaussian distribution, typically standard normal).
2. Construction of discrete factors using the `emdw` library.
3. Initialisation of cluster potentials and message passing as seen in Figure 3.2 to perform exact inference.
    - To avoid underflow, all factors were normalised after each update. This sacrificed some efficiency but significantly improved numerical stability.
4. Parameter update step (M-step).
5. Likelihood calculation and convergence check using the relative tolerance criterion:

$$\frac{|\ell(\Theta)^{\text{new}} - \ell(\Theta)^{\text{old}}|}{\ell(\Theta)^{\text{old}}} < \epsilon$$

   where the maximum number of iterations was capped at 100, whilst the threshold value was chosen to be $\epsilon = 10^{-4}$.

modelSelection

The `modelSelection` function evaluated different values of the hyperparameter m:

1. For each candidate m, the model was run with 10 random restarts.
2. The best run (highest log-likelihood) was retained.
3. Model fit metrics (AIC, BIC, and maximum log-likelihood) were recorded for each m.
4. Results were exported to CSV files for analysis with Python.

3.4.4. Model Selection & Output

The final model outputs were exported in two forms:

- Posterior decoding using the Maximum Posterior Marginal (MPM) rule.
- State sequence decoding using the Viterbi algorithm.

Both outputs were written to CSV files by the `C++`, then processed and visualised in Python.

Chapter 4

Results

4.1. Model Decoding

foo

4.1.1. Viterbi

foo1

4.1.2. MPM Rule

foo2

Chapter 5

Body

5.1. Model Selection

5.2. Meditating a little bit more on model output

5.2.1. What I have been doing

We are computing the pointwise marginal MAP, often called the Maximum Posterior Marginal (MPM) rule. For each time t, we pick:

$$\hat{s}_t = \underset{s}{\text{argmax}} \;\; p(S_t = s \mid A_{1:T}, \Theta)$$

The MPM picks the most likely state at each time independently — which can lead to an impossible or very unlikely global sequence (eg. $S_t \equiv$ Very Wet, then $S_{t+1} \equiv$ Very Dry). It maximizes expected per-time classification accuracy, but it does not maximize the joint posterior probability of the entire sequence.

5.3. Viterbi Algorithm

The paper recommends the Viterbi algorithm, which finds

$$\mathbf{s}^* = \underset{\mathbf{S}_{1:T}}{\text{argmax}} \;\; p(\mathbf{S}_{1:T} \mid A_{1:T}, \Theta)$$

ie. The single state sequence with the highest joint posterior probability. That sequence respects transitions and is temporally coherent.

Okay, Just a little reminder, we have used the LBU paired with EM. I spoke to my professor and he mentioned that because of the model structure, we are actually constructing a Junction Tree meaning we get exact inference. Additionally, because of how the junction tree we start from the leaf nodes, the formulation of using the JTREE vs Forward-Backward is actually the exact same (Check math behind this...).

Anyway, we have our model output now $p(\mathbf{S}_{1:T} \mid A_{1:T}, \Theta)$ which is an exact measure. How do i know get the output of my model. The paper I am implementing says this:

"With the estimated optimal DNBC parameters, the most probable path of the latent drought state that maximizes P(A— ·) together with the probability of each state at every time step can be obtained using the Viterbi algorithm (Rabiner 1989)." Right now I am simply taking the maximum confidence for each $p(S_t \mid A_{1:T}, \Theta)$. This is most probably wrong. What must I do, explain to me what i must do and why what im doing is wrong (if it is wrong.)

Chapter 6

Summary and Conclusion

Bibliography

[1] J. Tyndall, "Global drought outlook — oecd," Jun 2025. [Online]. Available: https://www.oecd.org/en/publications/global-drought-outlook_d492583a-en.html

[2] S. Gebrechorkos, J. Sheffield, S. Vicente-Serrano, C. Funk, D. Miralles, J. Peng, E. Dyer, J. Talib, H. Beck, M. Singer, and S. Dadson, "Warming accelerates global drought severity," *Nature*, vol. 642, no. 8068, pp. 628–635, 6 2025.

[3] L. Chen, P. Brun, P. Buri, S. Fatichi, A. Gessler, M. Mccarthy, F. Pellicciotti, B. Stocker, and D. Karger, "Global increase in the occurrence and impact of multiyear droughts," *Science*, vol. 387, no. 6731, pp. 278–284, 1 2025.

[4] A. Olagunju, G. Thondhlana, J. S. Chilima, A. Sène-Harper, W. N. Compaoré, and E. Ohiozebau, "Water governance research in africa: progress, challenges and an agenda for research and action," *Water International*, vol. 44, no. 4, pp. 382–407, 2019. [Online]. Available: https://doi.org/10.1080/02508060.2019.1594576

[5] B. Shiferaw, K. Tesfaye, M. Kassie, T. Abate, B. Prasanna, and A. Menkir, "Managing vulnerability to drought and enhancing livelihood resilience in sub-saharan africa: Technological, institutional and policy options," *Weather and Climate Extremes*, vol. 3, pp. 67–79, 2014, high Level Meeting on National Drought Policy. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212094714000280

[6] C. Botai, J. Botai, J. De Wit, K. Ncongwane, and A. Adeola, "Drought characteristics over the western cape province, south africa," *Water*, vol. 9, no. 11, p. 876, 11 2017.

[7] I. B. Oluwatayo and T. M. Braide, "Socioeconomic determinants of households' vulnerability to drought in western cape, south africa," *Sustainability*, vol. 14, no. 13, 2022. [Online]. Available: https://www.mdpi.com/2071-1050/14/13/7582

[8] M.-A. Baudoin, C. Vogel, K. Nortje, and M. Naik, "Living with drought in south africa: lessons learnt from the recent el niño drought period," *International Journal of Disaster Risk Reduction*, vol. 23, pp. 128–137, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212420917300985

[9] P. M. Sousa, R. C. Blamey, C. J. C. Reason, A. M. Ramos, and R. M. Trigo, "The 'day zero' cape town drought and the poleward migration of moisture corridors," *Environmental Research Letters*, vol. 13, no. 12, p. 124025, dec 2018. [Online]. Available: https://dx.doi.org/10.1088/1748-9326/aaebc7

[10] R. C. Odoulami, P. Wolski, and M. New, "A som-based analysis of the drivers of the 2015–2017 western cape drought in south africa," *International Journal of Climatology*, vol. 41, no. S1, pp. E1518–E1530, 2021. [Online]. Available: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.6785

[11] L. S. Joubert and G. Ziervogel, *Day zero: One city's response to a record-breaking drought*. University of Cape Town, 2019.

[12] P. A. N. Babajide Olusola Sanwo-Olu, K. S. Michael Danquah, R. Calland, L. S. Brahima Sangafowa Coulibaly, L. S. Vera Songwe, and F. G. Ahmadou Aly Mbaye, "Cape town: Lessons from managing water scarcity," May 2023. [Online]. Available: https://www.brookings.edu/articles/cape-town-lessons-from-managing-water-scarcity/

[13] D. C. Edossa, Y. E. Woyessa, and W. A. Welderufael, "Analysis of droughts in the central region of south africa and their association with sst anomalies," *International Journal of Atmospheric Sciences*, vol. 2014, no. 1, p. 508953, 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2014/508953

[14] B. Lloyd-Hughes, "The impracticality of a universal drought definition," *Theoretical and Applied Climatology*, vol. 117, 10 2013.

[15] D. Wilhite and M. Glantz, "Understanding: the drought phenomenon: The role of definitions," *Water International - WATER INT*, vol. 10, pp. 111–120, 01 1985.

[16] T. B. McKee, N. J. Doesken, J. Kleist *et al.*, "The relationship of drought frequency and duration to time scales," in *Proceedings of the 8th Conference on Applied Climatology*, vol. 17, no. 22. California, 1993, pp. 179–183.

[17] H. Douville, K. Raghavan, J. Renwick, R. P. Allan, P. A. Arias, M. Barlow, R. Cerezo-Mota, A. Cherchi, T. Gan, J. Gergis *et al.*, "Water cycle changes," 2021.

[18] I. Nalbantis and G. Tsakiris, "Assessment of hydrological drought revisited," *Water resources management*, vol. 23, no. 5, pp. 881–897, 2009.

[19] A. Van Loon, "Hydrological drought explained," *Wiley Interdisciplinary Reviews: Water*, vol. 2, 04 2015.

[20] J. Judith, R. Tamilselvi, M. P. Beham, S. Lakshmi, A. Panthakkan, S. A. Mansoori, and H. A. Ahmad, "Remote sensing based crop health classification using ndvi and fully connected neural networks," *arXiv preprint arXiv:2504.10522*, 2025.

[21] G. Maracchi, *Agricultural Drought — A Practical Approach to Definition, Assessment and Mitigation Strategies*. Dordrecht: Springer Netherlands, 2000, pp. 63–75. [Online]. Available: https://doi.org/10.1007/978-94-015-9472-1_5

[22] D. Ji, X. Li, Y. Niu, S. Chen, Y. Huang, and S. Zhou, "Response strategies to socio-economic drought: An evaluation of drought resistance capacity from a reservoir operation perspective," *Water*, vol. 17, no. 7, 2025. [Online]. Available: https://www.mdpi.com/2073-4441/17/7/1002

[23] T. Wang, X. Tu, V. P. Singh, X. Chen, K. Lin, R. Lai, and Z. Zhou, "Socioeconomic drought analysis by standardized water supply and demand index under changing environment," *Journal of Cleaner Production*, vol. 347, p. 131248, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959652622008794

[24] A. Mehran, O. Mazdiyasni, and A. AghaKouchak, "A hybrid framework for assessing socioeconomic drought: Linking climate variability, local resilience, and demand," *Journal of Geophysical Research: Atmospheres*, vol. 120, no. 15, pp. 7520–7533, 2015. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JD023147

[25] F. M. Chivangulula, M. Amraoui, and M. G. Pereira, "The drought regime in southern africa: A systematic review," *Climate*, vol. 11, no. 7, 2023. [Online]. Available: https://www.mdpi.com/2225-1154/11/7/147

[26] H. Mulenga, M. Rouault, and C. Reason, "Dry summers over ne south africa and associated circulation anomalies," *Climate Research - CLIMATE RES*, vol. 25, pp. 29–41, 10 2003.

[27] [Online]. Available: https://www.drought.gov/what-is-drought/monitoring-drought

[28] Oct 2024. [Online]. Available: https://www.ncei.noaa.gov/news/making-drought-map

[29] [Online]. Available: https://droughtmonitor.unl.edu/About/WhatistheUSDM.aspx

[30] [Online]. Available: https://joint-research-centre.ec.europa.eu/european-and-global-drought-observatories/current-drought-situation-europe_en

[31] E. Esfahanian, A. P. Nejadhashemi, M. Abouali, U. Adhikari, Z. Zhang, F. Daneshvar, and M. R. Herman, "Development and evaluation of a comprehensive drought index," *Journal of environmental management*, vol. 185, pp. 31–43, 2017.

[32] M. B. Mukhawana, T. Kanyerere, and D. Kahler, "Review of in-situ and remote sensing-based indices and their applicability for integrated drought monitoring in south africa," *Water*, vol. 15, no. 2, 2023. [Online]. Available: https://www.mdpi.com/2073-4441/15/2/240

[33] H. Kim, D.-H. Park, J.-H. Ahn, and T.-W. Kim, "Development of a multiple-drought index for comprehensive drought risk assessment using a dynamic

naive bayesian classifier," *Water*, vol. 14, no. 9, 2022. [Online]. Available: https://www.mdpi.com/2073-4441/14/9/1516

[34] S. Chen, W. Muhammad, J.-H. Lee, and T.-W. Kim, "Assessment of probabilistic multi-index drought using a dynamic naive bayesian classifier," *Water Resources Management*, vol. 32, no. 13, pp. 4359–4374, 8 2018.

[35] B. Poudel, D. Dahal, S. Shrestha, R. Sewa, and A. Kalra, "Developing a composite drought indicator using pca integration of chirps rainfall, temperature, and vegetation health products for agricultural drought monitoring in new mexico," *Atmosphere*, vol. 16, no. 7, 2025. [Online]. Available: https://www.mdpi.com/2073-4433/16/7/818

[36] S. L. Lauritzen and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 50, no. 2, pp. 157–194, 1988.

[37] J. Binder, K. Murphy, and S. Russell, "Space-efficient inference in dynamic probabilistic networks," *Bclr*, vol. 1, p. t1, 1997.

[38] "Forward–backward algorithm," Aug 2025. [Online]. Available: https://en.wikipedia.org/wiki/Forward%E2%80%93backward_algorithm

[39] H. Avilés-Arriaga, L. Sucar, C. Mendoza-Durán, and L. Pineda, "A comparison of dynamic naive bayesian classifiers and hidden markov models for gesture recognition," *Journal of applied research and technology*, vol. 9, pp. 81–102, 04 2011.

[40] E. Xing, "Junction tree algorithm and a case study of the hidden markov models," 2007. [Online]. Available: https://www.cs.cmu.edu/~epxing/Class/10708-07/Slides/lecture6-JT.pdf

[41] T. Moon, "The expectation-maximization algorithm," *Signal Processing Magazine, IEEE*, vol. 13, pp. 47 – 60, 12 1996.

[42] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models - Chapter A: Hidden Markov Models*, 3rd ed., 2025, online manuscript released August 24, 2025.

[43] E. Xing, "Lecture 6: Case studies: Hmm and crf," 2020. [Online]. Available: https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.cs.cmu.edu/~epxing/Class/10708-20/scribe/lec4_scribe.pdf&ved=2ahUKEwi9te3BtYKQAxVcQkEAHcQLKPkQFnoECBsQAQ&usg=AOvVaw1eG_6Kg3WNAg9dKdc1WOeV

[44] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

[45] S. Helske and J. Helske, "Mixture hidden markov models for sequence data: The seqhmm package in r," *Journal of statistical software*, vol. 88, 01 2019.

[46] Y.-C. Chen, "Lecture 9: Hidden markov model," https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=http://faculty.washington.edu/yenchic/18A_stat516/Lec9_HMM.pdf&ved=2ahUKEwig8Z36v4OQAxXGUUEAHX7aMIUQFnoECBYQAQ&usg=AOvVaw3VZuXe7Qh8Kc7F1G-H92uj, 2018.

Appendix A

Project Planning Schedule

This is an appendix.

Appendix B

Outcomes Compliance

This is another appendix.