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

Drought is a recurring and complex environmental challenge in South Africa, impacting agriculture, water security, and socio-economic stability. Existing monitoring systems often rely on individual indicators that capture only isolated dimensions of drought behaviour. This project develops a probabilistic, composite drought indicator using a Dynamic Naive Bayes Classifier (DNBC) to integrate three key indices: the Standardised Precipitation Index (SPI), Streamflow Drought Index (SDI), and Normalised Difference Vegetation Index (NDVI). Together, these represent meteorological, hydrological, and agricultural dimensions of drought.

The model was implemented for the period 1981–2019 in the Western Cape using open-source rainfall, streamflow, and remote-sensing datasets. Each index was computed, preprocessed, and discretised to serve as observed variables, while latent drought states were inferred through parameter estimation. This was done using the Expectation–Maximisation (EM) algorithm with Junction Tree (JT) inference. Model selection was guided by information criteria which identified six latent drought states representing varying degrees of dryness and wetness.

Results show that the DNBC successfully identified major historical drought periods in South Africa, demonstrating comparable, and in some respects improved, performance relative to the individual input indices. Quantitatively, the DNBC achieved the highest F1-score, indicating a stronger balance between sensitivity and precision under highly imbalanced data conditions. Although the model exhibited oscillatory behaviour not reflective of drought dynamics and produced false alarms, it successfully captured the abstract dimensions of drought.

The study concludes that a probabilistic approach such as the DNBC offers a valuable foundation for operational drought monitoring, particularly due to its ability to express uncertainty in classification. Future work should focus on improving data quality, incorporating continuous variables, and extending the framework to additional drought indicators to promote stability and robustness.

Afrikaans

Droogte is ’n herhalende en komplekse omgewingsuitdaging in Suid-Afrika wat landbou, watersekerheid en sosio-ekonomiese stabiliteit beïnvloed. Bestaande moniteringstelsels

steun dikwels op individuele aanwysers wat slegs beperkte aspekte van droogtegedrag vasvang. Hierdie projek ontwikkel 'n waarskynlikheidsgebaseerde, saamgestelde droogte-aanwyser deur middel van 'n Dynamic Naive Bayes Classifier (DNBC) om drie sleutelindekse te integreer: die Standardised Precipitation Index (SPI), Streamflow Drought Index (SDI) en Normalised Difference Vegetation Index (NDVI). Saam verteenwoordig hierdie indekse die meteorologiese, hidrologiese en landboukundige dimensies van droogte.

Die model is vir die tydperk 1981–2019 in die Wes-Kaap geïmplementeer met behulp van oopbron reënval-, stroomvloeien- en afstandwaarnemingsdatastelle. Elke indeks is bereken, voorafverwerk en gediskretiseer om as waargenome veranderlikes te dien, terwyl latente droogtetoestande afgelei is deur middel van parameterberaming. Dit is uitgevoer met die Expectation–Maximisation (EM) algoritme saam met Junction Tree (JT) inferensie. Modelseleksie is gelei deur inligtingsteoretiese kriteria wat ses latente toestande geïdentifiseer het wat verskillende vlakke van droogheid en natheid voorstel.

Die resultate toon dat die DNBC suksesvol groot historiese droogteperiodes in Suid-Afrika geïdentifiseer het en vergelykbare, en in sekere opsigte verbeterde, prestasie gelewer het relatief tot die individuele insetindekse. Kwantitatief het die DNBC die hoogste F1-telling behaal, wat 'n beter balans tussen sensitiwiteit en presisie aandui in 'n sterk ongebalanseerde datastel. Alhoewel die model ossillerende gedrag getoon het wat nie droogtedinamika akkuraat weerspieël nie en valse alarms gegenereer het, het dit die abstrakte dimensies van droogte effektief vasgevang.

Die studie kom tot die gevolgtrekking dat 'n waarskynlikheidsbenadering soos die DNBC 'n waardevolle grondslag bied vir operasionele droogtemonitering, veral vanweë sy vermoë om onsekerheid in klassifikasie uit te druk. Toekomstige werk behoort te fokus op die verbetering van datagehalte, die inkorporering van deurlopende veranderlikes, en die uitbreiding van die raamwerk na addisionele droogte-aanwysers om stabiliteit en robuustheid te bevorder.

Contents

List of Figures

List of Tables

Nomenclature

Variables and functions

T	Total number of time steps
t	A single time step index
N	Total number of input variables
n	A single input variable index
S_t	Random variable for the latent drought state at time t
$A_t^{(n)}$	Observed value of the n-th input variable at time t
Θ	The set of all model parameters
π_i	Initial state probability: $p(S_1 = i)$
$a_{i,j}$	Transition probability from latent state i to j: $p(S_t = j \mid S_{t-1} = i)$
$b_i^{(n)}(j)$	Emission probability: $p(A_t^{(n)} = j \mid S_t = i)$
m	Number of possible latent drought states (cardinality of S_t)
C_n	Number of discrete values for the n-th input variable $A_t^{(n)}$
$\mathbf{A}_t$	Set of all input variables at time t: $\{A_t^{(1)}, \ldots, A_t^{(N)}\}$
$A_{1:T}$	Set of all input variables across all time steps: $\{\mathbf{A}_1, \ldots, \mathbf{A}_T\}$
$\mathbf{S}_{1:T}$	Set of all latent state variables across all time steps: $\{S_1, \ldots, S_T\}$
$\mathcal{H}$	The set of all latent states in the DNBC
$\mathcal{D}$	The set of all observed data in the DNBC
$\ell(\Theta)$	The log-likelihood function of the model
$L(\Theta)$	The maximised log-likelihood of the model
$\psi_i(\mathbf{X})$	Cluster potential of cluster i over variables $\mathbf{X}$ in the junction tree
$\delta_{i \to j}(\mathbf{S})$	Message from cluster i to j over separator set $\mathbf{S}$ in the junction tree
$p(X)$	Probability mass function of the discrete random variable X
ε	A small positive number (tolerance/threshold)

Acronyms and abbreviations

DNBC	Dynamic Naive Bayes Classifier
HMM	Hidden Markov Model
JT	Junction Tree
SPI	Standard Precipitation Index
SDI	Streamflow Drought Index
NDVI	Normalized Difference Vegetation Index
EM	Expectation–Maximisation
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
DWS	Department of Water and Sanitation
SAWS	South African Weather Service
RV	Random Variable
MPM	Maximum Posterior Marginal

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, which are already under threat due to poor infrastructure and inequitable management. These two issues are prevalent in South Africa and 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) or the Standardized Precipitation-Evapotranspiration Index (SPEI). These indices compare current precipitation levels to long-term historical averages for a specific region [16–19].

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 and is measured using indices such as the Streamflow Drought Index (SDI) or the Standardized Streamflow Index (SSI), which are metrics derived from river monitoring [19–22].

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) [23]. Another common index to use for this aspect of drought is the Evaporative Stress Index (ESI) which quantifies anomalies in evapotranspiration. 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 [19, 24–26].

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 [27]. 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 [19, 28, 29].

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 [30, 31].

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 [32].

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 [33].

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 [34–36].

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 [30, 37].

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 [38, 39].

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

This project develops and evaluates a probabilistic framework for drought monitoring in South Africa using a DNBC. The model integrates three complementary drought indicators, that being the SPI, SDI and NDVI, to construct a composite drought indicator that encapsulates meteorological, hydrological, and agricultural dimensions of drought within a unified, probabilistic structure.

The study spans the period 1981–2019 in the Western Cape and relies exclusively on open-source datasets. Substantial effort was spent on acquisition, cleaning, and preprocessing the of the data to ensure quality. The SPI and SDI were derived directly from these datasets, while NDVI was incorporated as a satellite-based measure of vegetation stress.

The DNBC was designed with discrete latent variables representing drought states and observed nodes corresponding to the input indices. Parameter estimation was performed using the Expectation–Maximisation (EM) algorithm in conjunction with Junction Tree (JT) inference for efficient computation of marginal and joint probabilities. Model selection was guided by the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and the maximised log-likelihood.

Model performance was assessed both qualitatively and quantitatively. Temporal plots revealed that the DNBC effectively identified major historical drought events in South Africa. Although the model exhibited short-term oscillations and false alarms, its predictive accuracy was found to be comparable to, and in some aspects superior to, the individual drought indices by achieving the highest F1-score among all other indices. This indicates that the model successfully integrates the input drought indices to capture complex drought dynamics that no single index can.

1.5. Scope

While the ultimate goal of drought monitoring is to capture each dimension of drought, including the socio-economic aspect, this project deliberately restricts its scope to the meteorological, hydrological, and agricultural domains. The exclusion is primarily due to the considerable complexity involved in quantifying socio-economic indicators in the South African context. For existing methods of quantifying this dimension of drought, the limited availability of reliable, open-access data makes it near impossible.

Similarly, the project relies on openly available precipitation data obtained from the University of Cape Town (UCT) rather than proprietary datasets from the South African Weather Service (SAWS). While the coverage of this dataset makes it well-suited for this project's purpose, its validity is uncertain when compared to the official SAWS records.

Finally, it will also be noted that, fundamentally, the DNBC itself has several naive assumptions which will be discussed later. Although these assumptions are naive, the purpose of this study is to evaluate the effectiveness of this framework as a drought monitoring tool in the South African context.

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 [38].

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 [39].

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 [37].

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 [40].

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 [30, 37–39].

Fix only the last sentence, leave everyhting else as is

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 [41]). The dataset provides rainfall values at station level. This offers a high degree of 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 [42]). These daily records were averaged to create monthly records, which was then used to calculate the desired index.

To obtain the NDVI, the NOAA Climate Data Record (CDR) of AVHRR Normalised Difference Vegetation Index (NDVI), Version 5 was used (Dataset [43]). 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.

It is important to note that the overlapping period of the available data, and thus the scope for the model output spans from 1981 to 2019.

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 collected data was transformed into drought indices using well-known methodologies. As mentioned, each index captures different dimensions of drought conditions. Together they provide a more comprehensive view than any single measure.
Below is a brief overview of each index and its mathematical formulation. Full derivations and discussions are available in the cited references.

3.2.1. Standardised Precipitation Index (SPI)

The SPI is based solely on precipitation and measures anomalies relative to the long-term probability distribution at a given location and timescale. The calculation involves fitting a long-term precipitation record to a gamma distribution, which is then transformed into a standard normal distribution [16].

The SPI can be computed for different time scales (e.g., 1, 3, 6, or 12 months). This study employs the 3-month timescale (SPI-3), which is widely used for monitoring. Negative SPI values indicate dry conditions, with established thresholds categorising drought severity.

3.2.2. Streamflow Drought Index (SDI)

The SDI extends the framework of the SPI to characterise hydrological drought using streamflow volumes. Monthly streamflow observations are aggregated and standardized in a similar manner to the SPI calculation [20].

This method allows for the identification of periods with below-normal streamflow, indicating hydrological drought. Negative SDI values denote dry conditions, and the same severity thresholds used for the SPI are applied for categorisation.

3.2.3. Normalised Difference Vegetation Index (NDVI)

The NDVI is a remote-sensing indicator widely used to monitor vegetation health and stress, including agricultural drought. It is derived from the difference in reflectance between the near-infrared (NIR) and red spectral bands:

$$\mathrm{NDVI} = \frac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{RED}}}{\rho_{\mathrm{NIR}} + \rho_{\mathrm{RED}}},$$

where ρ_{NIR} and ρ_{RED} are the surface reflectances in their respective bands. Values range from -1 to $+1$, with higher values indicating healthy, active vegetation, while lower values reflect stressed or sparse vegetation [24].

3.2.4. Discretisation of Indices

The SPI, SDI, and NDVI are all continuous. However, the proposed model requires discrete inputs. Accordingly, each index was discretised into categorical bins based on thresholds widely used in literature. For example, SPI values are often classified into categories such as "extremely dry," "moderately dry," and "normal." This discretisation not only aids in model implementation but it also motivates interpretability. Table 3.1 below illustrates the bins used:

Table 3.1: Discretisation thresholds for drought indices.

Category	SPI / SDI
Severe Drought	≤ -1.5
Moderate Drought	$-1.5 < x \leq -0.5$
Normal	$-0.5 < x < 0.5$
Moderate Wet	$0.5 \leq x < 1.5$
Severe Wet	≥ 1.5

Category	NDVI
Bare soil / water	$-1 < x < 0.1$
Sparse vegetation	$0.1 \leq x < 0.2$
Moderate vegetation	$0.2 < x < 0.4$
Dense vegetation	$0.4 \leq x < 0.6$
High density vegetation	$0.6 \leq x < 1$

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.

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.

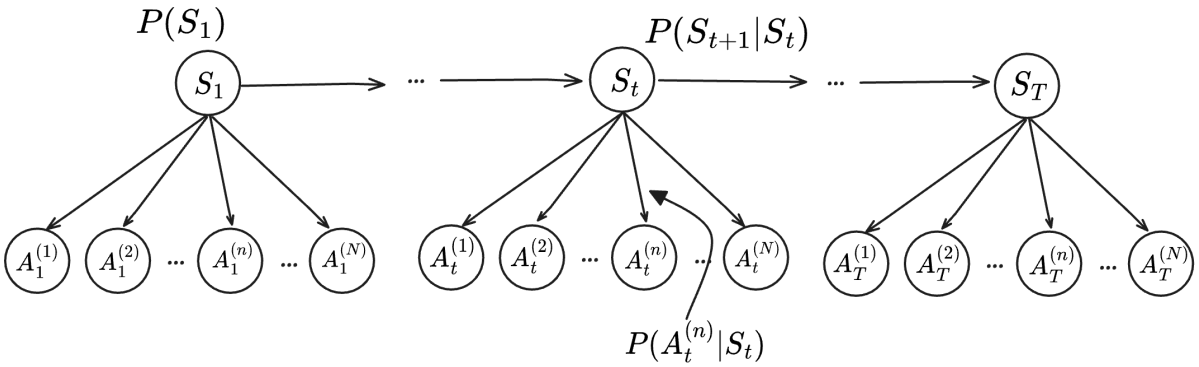

Figure 3.1: The Dynamic Naive Bayes Classifier (DNBC) can be represented as a Bayesian network unfolding over time. At each time step t, a latent drought state S_t is modelled as a discrete random variable that governs the latent structure, while the observed input variables $\mathbf{A}_t = \{A_t^{(1)},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

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 state 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	$a_{1,2}$
$\vdots$	$\vdots$	$\vdots$
1	m	$a_{1,m}$
2	1	$a_{2,1}$
2	2	$a_{2,2}$
$\vdots$	$\vdots$	$\vdots$
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 the 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 input variable, given that the system is in a specific hidden state. These parameters encode how the drought indicators behave under each latent drought 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 that its corresponding hidden drought state is equal to i

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

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 [44].
Figure 3.2 illustrates the JT structure associated with the DNBC.

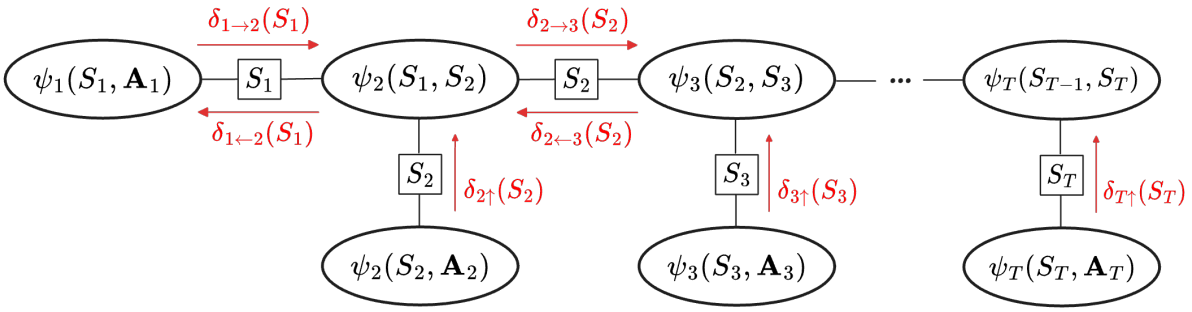

Figure 3.2: Junction Tree representation of the DNBC. Each cluster groups together latent state variables and observed attributes, with sepsets defined along the edges. Messages are propagated through the tree to perform exact inference.

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.

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

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

Message Passing

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

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:

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

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

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* [45–47].
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 [48].

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 [39, 47–49].

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 [50]. 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 [51, 52]:

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

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

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

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 [53]. 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 [54], 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 [54, 55] 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 con-

struction 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.
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.

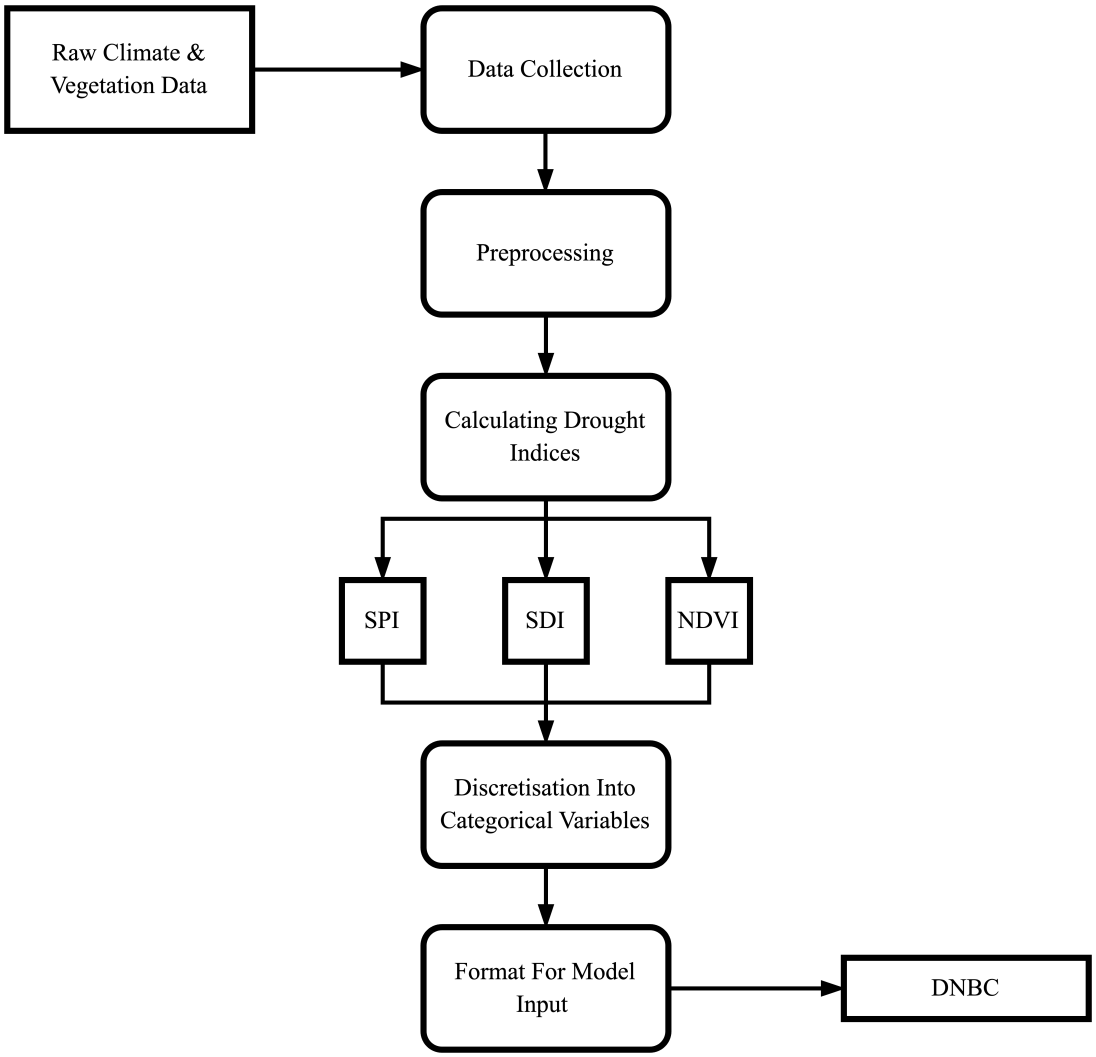

Figure 3.3: Data pipeline from raw climate and vegetation data to discretised indices used as DNBC inputs. The pipeline consists of five stages: data collection, preprocessing, index calculation, discretisation, and input formatting.

- 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:

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

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 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.

1. Maximum Posterior Marginal (MPM) rule

The MPM rule involves computing the point-wise marginal for latent drought state S_t:

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

This is easily obtained using the results from Section 3.3.2.

Conceptually, this rule will pick the most likely state, with a confidence attached, at each time step independently. This is particularly useful if the goal is signal labeling. It should be noted that this rule often leads to an unlikely or even impossible state sequence (eg., $S_t \equiv$ Very Wet $\parallel$ $S_{t+1} \equiv$ Very Dry) and thus is only sensible to use when temporal consistency can be ignored.

2. Viterbi Algorithm

On the other hand, the Viterbi algorithm produces a temporally coherent sequence that respects the state transition dynamics. Mathematically, the Viterbi decoding produces the single most probable joint state sequence:

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

This decoding provides a more realistic representation of drought progression [56].

Output Format

Both outputs were exported to CSV files by the `C++` implementation. Post-processing, visualisation, and evaluation of these outputs were performed in Python, enabling comparison with the original input indices and facilitating analysis.

Chapter 4

Results

The following chapter presents the results obtained from the DNBC. The analysis begins with model selection then moves to qualitative assessment of state outputs, and finally to a quantitative evaluation of the model's classification performance against the input drought indices used as a benchmark.

4.1. Model Selection & State Definition

The optimal number of latent states was found to be 6. Selection criteria stipulated in Section 3.3.4 was met on all fronts as Figure 4.1 shows a minimum value of both BIC and AIC are found at $m = 6$. Additionally, the graph shows a pronounced jump in log-likelihood improvement at this point. While the log-likelihood continues to increase for larger m, the marginal gains are smaller for m values greater than 6 which may suggest that those models have too many parameters.

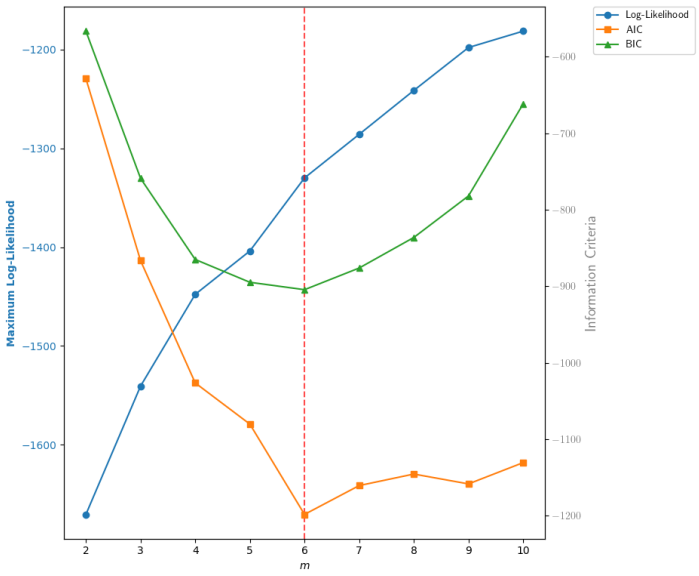

Figure 4.1: Plot of AIC, BIC and maximum log-likelihood across m (number of hidden states)

Thus, subsequent analyses use $m = 6$, corresponding to the following classification of

latent states:

$$1 : (S3D) \equiv \text{Extreme Drought},$$
$$2 : (S2D) \equiv \text{Moderate Drought},$$
$$3 : (S1D) \equiv \text{Mild Drought},$$
$$4 : (S1W) \equiv \text{Mild Wet},$$
$$5 : (S2W) \equiv \text{Moderate Wet},$$
$$6 : (S3W) \equiv \text{Extreme Wet}.$$

4.2. Model Behaviour

4.2.1. Latent-State Sequence Output

Figure 4.2 presents the Viterbi-decoded state sequence from 1981 to 2019, along with the known historical drought periods identified in literature that overlap with the model's output as shaded regions, that being 1983–1984, 1991–1992, 1994–1995, 2014–2016, and 2017–2018

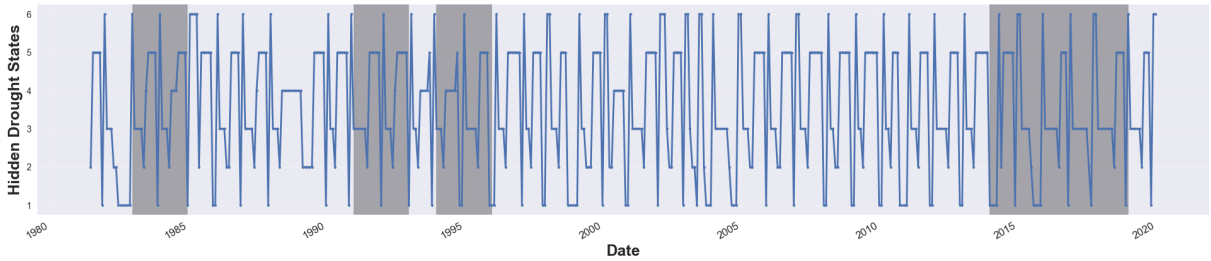

Figure 4.2: Viterbi-decoded state sequence of DNBC with known drought periods as shaded regions

This plot shows that its performance in identifying historical drought periods is mixed, showing some promising signals but also significant weaknesses. There is a clear tendency for the model to enter higher states during the shaded periods which indicates that the model is successfully characterising these events.

However, the model displays extreme volatility and oscillatory behaviour. The state sequence frequently fluctuates between the highest and lowest states over short periods. This is not reflective of real world drought dynamics which persists over months or years. Furthermore, the model identifies many periods as "Extreme Drought" outside of the documented historical events. These occurrences are clear false positives, which greatly hinder the model's reliability for drought monitoring.

4.2.2. Model Confidence and Input Comparison

To better visualise the relationships between the input indices and the model output, Figure 4.3 shows the SPI, SDI, and NDVI time series alongside the DNBC output. At each monthly time step, the DNBC's output has a confidence attached which is represented by the height of the vertical bar. This confidence is derived from the MPM Rule probabilities while the colour represents the classification which comes from the Viterbi output.

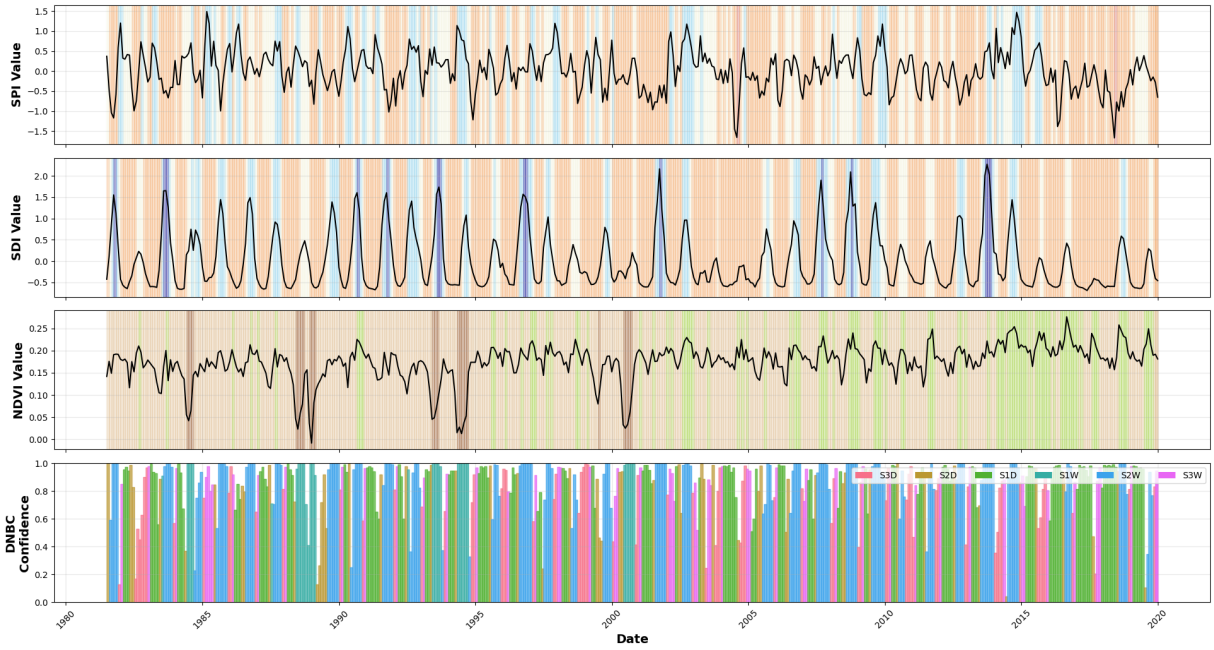

Figure 4.3: Drought classifications for the period 1981-2019 using SPI, SDI, NDVI, and DNBC (the vertical, coloured bars represent different drought states, while their height indicates the confidence in classification. The black lines plot the continuous values of SPI, SDI and NDVI)

This graph on the other hand shows that the input indices themselves are inherently noisy and oscillatory, reflecting the high variability of environmental conditions and in turn the model output. This volatility could also contribute to the model's uncertain classifications that are present in the plot. Nonetheless, this probabilistic approach allows these uncertainties to be explicitly shown, which these standard indices lack.

4.3. Quantitative Evaluation

A quantitative evaluation was conducted by treating known drought events as a binary classification problem. For the period 1981–2019, each month was classified as either a drought, considered the positive class, or non-drought, the negative class. The predictions from the model and each input index were then compared against these historical classifications. The resulting confusion matrices are presented in Figure 4.4. Using these matrices, performance metrics were calculated and are summarised in Table 4.1.

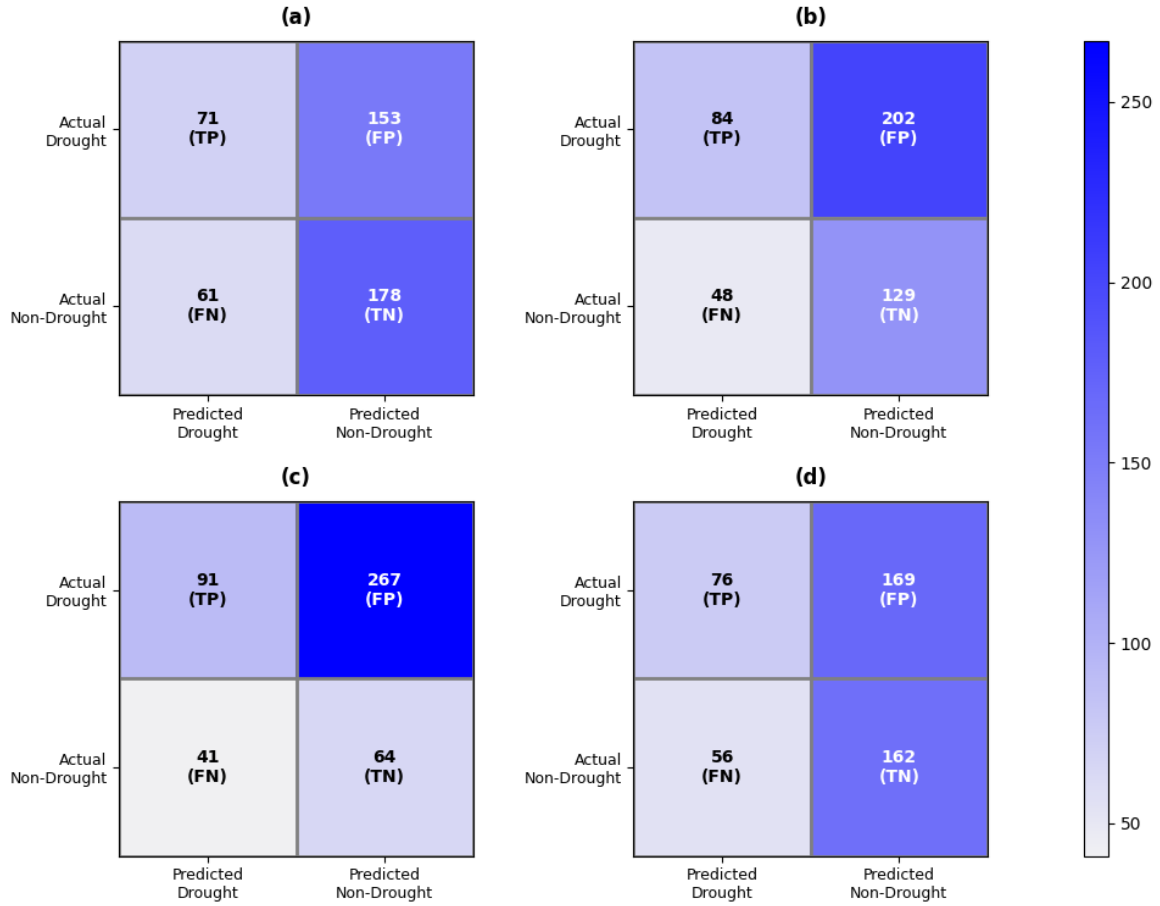

Figure 4.4: Confusion matrices for classifying known drought states using (a) SPI, (b) SDI, (c) NDVI, (d) DNBC.

Table 4.1: Performance comparison of three input indices (SPI, SDI, NDVI) and the DNBC model in classifying drought events. The models are evaluated using four standard metrics: Recall, Accuracy, Precision, and F1 Score.

Indicator	Recall (%)	Accuracy (%)	Precision (%)	F1 Score (%)
SPI	53.79	53.78	31.70	39.89
SDI	63.64	46.00	29.37	40.19
NDVI	68.94	33.48	25.42	37.14
DNBC	57.58	51.40	31.02	40.32

4.3.1. Performance Metrics and Interpretation

In this context, the metrics of Recall and Precision evaluate two distinct and critical aspects of model performance. Recall measures the model's ability to correctly identify actual drought events. Thus, high Recall is crucial for a warning system, as it means fewer droughts are missed. Conversely, Precision measures the reliability of the model's drought alarms as it indicates that when the model predicts a drought, it is likely to be correct. In practice, scoring high in Precision directly reduces false alarms and associated costs.

The F1-score is particularly informative in highly imbalanced datasets such as this one where drought events form a small fraction of the total. Additionally, accuracy can be misleading in this context, since a model that always predicts "no drought" could achieve high accuracy simply due to class imbalance. Hence, F1-score is the most appropriate performance metric for this application.

The following observations can be made from the results:

- **NDVI:** Exhibits high Recall (68.9%) but low Precision (25.4%), indicating that it frequently identifies drought conditions, including many false alarms. This could be due to NDVI's sensitivity to the agricultural aspect of drought, which can both lag or persist beyond meteorological drought. This will lead to overestimation.
- **SPI:** Shows lower Recall (53.8%) and moderate Precision (31.7%), suggesting it provides a more conservative estimate of drought.
- **SDI:** Achieves Recall (63.6%) similar to NDVI but slightly better Precision (29.4%). SDI therefore offers a slight improvement to the NDVI, as shown by the increased F1-score (40.2%)
- **DNBC Output (Viterbi):** Marginally achieves the highest F1-score (40.3%), with Recall (57.6%) and Precision (31.0%). This indicates that the DNBC achieves a more stable trade-off between false alarms and missed events, suggesting that it captures the shared structure among the indices.

Chapter 5

Summary and Conclusion

This project sought to develop and evaluate a principled, probabilistic approach for drought monitoring in the South African context using a Dynamic Naive Bayes Classifier (DNBC). The model combines three key drought indicators, the Standardised Precipitation Index (SPI), Streamflow Drought Index (SDI), and Normalised Difference Vegetation Index (NDVI), to construct a composite drought indicator that captures the complexity of meteorological, hydrological, and agricultural dimensions of drought.

The approach was implemented over the period of 1981–2019 with the study area being the Western Cape using open source datasets. The DNBC was formulated with discrete random variables representing latent drought states and observed input indices. Parameter estimation was performed using the Expectation–Maximisation (EM) algorithm paired with Junction Tree (JT) inference, and model selection was guided by AIC, BIC, and maximised log-likelihood criteria.

The model was evaluated both qualitatively and quantitatively, with plots revealing that the DNBC successfully identified the known drought periods in South Africa. However, not only did the model output exhibit oscillatory behaviour within relatively short time periods, showing a false representation of actual drought dynamics, it also produced many false alarms for its classifications. Nonetheless, the DNBC's performance was found to be comparable to, and in some respects better than, the individual indices as it achieved the highest F1-score among all evaluated methods. This suggests that the composite indicator captures abstract information across the different drought dimensions.

Reflection on Objectives

The objectives set out at the beginning of this work were to develop a composite drought indicator using a DNBC, to evaluate its performance relative to established indices, and to assess its applicability within the South African context. Each of these objectives was achieved. The DNBC framework was successfully designed and implemented, its performance compared against established drought indices, and its results were obtained despite the unique data and climatic challenges of South Africa. Overall, the findings indicate that the DNBC-based composite indicator performs at least comparably to existing

indices, and arguably better.

Future Work and Recommendations

Although the DNBC demonstrated promising results, several paths exist for further improvement and exploration:

- **Continuous Inputs:** This implementation discretised all input indices. Extending the DNBC to handle continuous random variables, for example via Gaussian or hybrid emission distributions, could preserve more information and reduce output variability.
- **Enhanced Data Quality:** Data reliability and consistency remain a significant limitation in the South African context. Access to higher quality rainfall, streamflow, and remote-sensing data would likely improve both the model's calibration and generalisation.
- **Expanded Input Set:** Future models could integrate additional indices such as soil moisture, evapotranspiration, or temperature-based indicators to better capture multi-dimensional drought processes.
- **Alternative Methods:** Exploring other probabilistic and machine learning approaches such as Random Forests, Support Vector Machines or even deep learning methods may have a greater capacity to capture the complexity of drought.

In summary, this work demonstrates that probabilistic graphical models, specifically the DNBC, represent a viable approach to drought characterisation in data-limited environments. While challenges remain, particularly regarding data availability and noise in model output, the results show that integrating multiple drought dimensions using a principled and probabilistic approach yields both interpretive and operational value. With further refinement and expanded datasets, this approach could form the foundation of a robust drought monitoring system for South Africa.

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] S. M. Vicente-Serrano, S. Beguería, and J. I. Ló pez Moreno, "A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index," *Journal of climate*, vol. 23, no. 7, pp. 1696–1718, 2010.

[19] M. D. Svoboda, B. A. Fuchs *et al.*, *Handbook of drought indicators and indices*. World Meteorological Organization Geneva, Switzerland, 2016, vol. 2.

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

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

[22] S. M. Vicente-Serrano, J. I. López-Moreno, S. Beguer ía, J. Lorenzo-Lacruz, C. Azorin-Molina, and E. Morán-Tejeda, "Accurate computation of a streamflow drought index," *Journal of Hydrologic Engineering*, vol. 17, no. 2, pp. 318–332, 2012.

[23] 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.

[24] C. J. Tucker, "Red and photographic infrared linear combinations for monitoring vegetation," *Remote Sensing of Environment*, vol. 8, no. 2, pp. 127–150, 1979. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0034425779900130

[25] 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

[26] M. C. Anderson, C. A. Zolin, P. C. Sentelhas, C. R. Hain, K. Semmens, M. T. Yılmaz, F. Gao, J. A. Otkin, and R. Tetrault, "The evaporative stress index as an indicator of agricultural drought in brazil: An assessment based on crop yield impacts," 2016.

[27] 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

[28] 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

[29] 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

[30] 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

[31] 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.

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

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

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

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

[36] 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.

[37] 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

[38] 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

[39] 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.

[40] 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

[41] S. Conradie, B. Hewitson, and P. Wolski, "Winter rainfall zone 2019 station rainfall dataset," Sep 2021. [Online]. Available: https://zivahub.uct.ac.za/articles/dataset/Winter_Rainfall_Zone_2019_station_rainfall_dataset/16453452

[42] May 2011. [Online]. Available: https://www.dws.gov.za/hydrology/Verified/hymain.aspx

[43] E. Vermote, N. Cdr, and Program, "Noaa climate data record (cdr) of avhrr normalized difference vegetation index (ndvi), version 5," 2019.

[44] 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.

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

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

[47] 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.

[48] 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

[49] "Baum–welch algorithm," Aug 2025. [Online]. Available: https://en.wikipedia.org/wiki/Baum%E2%80%93Welch_algorithm

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

[51] 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.

[52] 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

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

[54] 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.

[55] 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.

[56] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 2002.

Appendix A

Project Planning Schedule

This is an appendix.

Appendix B

Outcomes Compliance

This section outlines how the required Engineering Council of South Africa (ECSA) Graduate Attributes (GAs) were achieved throughout this project, with reference to the relevant report sections.

GA 1: Problem Solving

The project addressed the complex problem of drought monitoring in South Africa by developing a composite drought indicator using a probabilistic framework. This required identifying limitations in existing single-index approaches and formulating a model that could integrate multiple data sources (Section 1). The problem was analytically framed in probabilistic terms through the use of a Dynamic Naive Bayes Classifier (DNBC), where latent drought states were inferred from observable indices (Section 3.3.1). The integration of time-dependent stochastic modelling with environmental indices demonstrates the author's ability to identify, analyse, and solve a complex, multidisciplinary problem.

GA 2: Application of Scientific and Engineering Knowledge

The work applied mathematical, statistical, and computational knowledge to design and implement the DNBC. This included understanding and utilising probabilistic models, Bayesian inference, and the Expectation–Maximisation (EM) algorithm (Section 3.3.1 & Section 3.3.3). Furthermore, hydrological, meteorological, and remote-sensing knowledge was applied to compute the Standardised Precipitation Index (SPI), Streamflow Drought Index (SDI), and Normalised Difference Vegetation Index (NDVI) (Section 3.2). The synthesis of these diverse scientific domains illustrates the application of fundamental engineering and scientific principles to solve a real-world environmental problem.

GA 3: Engineering Design

The project required the procedural and non-procedural design of a data-driven system for drought classification. The DNBC architecture, including its latent and observed variable structure, was conceptualised and implemented to model the probabilistic dependencies between drought-related indices (Section 3.3.1). The model design process involved iterative

refinement, guided by model selection criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) (Section 3.3.4). This process reflects a structured design methodology that balances theoretical soundness with practical data limitations.

GA 4: Investigations, Experiments and Data Analysis

Significant experimental investigation was performed throughout the project. This included data acquisition and preprocessing (Section 3.1), index computation and discretisation (Section 3.2), and model training and evaluation (Section 4). The author conducted quantitative analyses such as precision, recall, and F1-score calculations to evaluate model performance. The qualitative assessment of temporal drought patterns further supported the interpretation of results. Together, these demonstrate competence in designing and conducting investigations and drawing valid, data-driven conclusions.

GA 5: Engineering Methods, Skills and Tools, Including Information Technology

This project required extensive use of computational tools and programming to achieve its results. The DNBC and its associated algorithms (EM, Viterbi, and Junction Tree) were implemented from first principles in `C++` using the `emdw` library. Furthermore, data processing and visualisation were implemented using `Python` accompanied by libraries such as `NumPy`, `pandas`, and `matplotlib`. The use of probabilistic graphical model theory, statistical computing, and open-source tools highlights the author's proficiency in modern engineering methods and IT-based tools (Section 3.4).

GA 6: Professional and Technical Communication

The author engaged in weekly in-person meetings with their supervisor to discuss progress, challenges, and next steps, ensuring clear and professional communication throughout the project. This report itself serves as a demonstration of formal technical writing ability, integrating complex mathematical and engineering concepts in a structured and coherent manner. The final oral presentation and project open day will further demonstrate the author's ability to communicate technical findings effectively to both academic and professional audiences.

GA 8: Individual Work

The project was completed entirely by the author, including the research, model design, coding, analysis, and report writing. While guidance was provided by their supervisor, all

implementation and problem-solving were conducted independently. This demonstrates the author's ability to plan, manage, and execute complex engineering tasks independently (Sections 1–5).

GA 9: Independent Learning Ability

The project required extensive self-directed learning in several unfamiliar domains. The author independently studied advanced probabilistic models such as Hidden Markov Models (HMMs), DNBCs, and associated algorithms including the Forward–Backward and Baum–Welch algorithms (Section 3.3). Additionally, significant effort was spent understanding drought indices (SPI, SDI, NDVI), their derivation, and interpretation within the South African context (Section 3.2). This demonstrates a high level of independent learning ability and adaptability to new technical challenges.