UNIVERSITEIT•STELLENBOSCH•UNIVERSITY
jou kennisvennoot • your knowledge partner

Developing a Composite Drought Indicator Using Probabilistic Graphical Models

Coen Potgieter
25999656

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.E. van Daalen

November 2025

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 remains a persistent challenge in South Africa, affecting agriculture, water security, and economic stability. Conventional monitoring often relies on single indices that capture limited aspects of drought behaviour. This study develops a probabilistic composite drought indicator using a dynamic naive Bayes classifier (DNBC) to integrate the standardised precipitation index (SPI), streamflow drought index (SDI), and normalised difference vegetation index (NDVI), representing meteorological, hydrological, and agricultural dimensions respectively.

The model, implemented for 1981–2019 in the southwestern Cape, estimated latent drought states using the expectation–maximisation (EM) algorithm with junction tree (JT) inference. Model selection, guided by standard criterion, identified a six-state configuration. Outputs were extracted via the Viterbi algorithm and Maximum Posterior Marginal (MPM) rule.

Results show the DNBC accurately identified major historical droughts and achieved the highest F1-score among inputs, demonstrating reliable classification and balanced sensitivity, supporting its use for probabilistic drought monitoring.

Afrikaans

Droogte bly 'n volgehoue uitdaging in Suid-Afrika en beïnvloed landbou, watersekerheid en ekonomiese stabiliteit. Konvensionele monitering steun dikwels op enkele indekse wat slegs beperkte aspekte van droogtegedrag vasvang. Hierdie studie ontwikkel 'n waarskynlikheidsgebaseerde, saamgestelde droogte-aanwyser deur gebruik te maak van 'n dinamiese naïewe Bayes-klassifiseerder (DNBC) wat die gestandaardiseerde reënvalindeks (SPI), stroomvloei-droogte-indeks (SDI) en genormaliseerde verskilmplantindeks (NDVI) integreer, wat onderskeidelik die meteorologiese, hidrologiese en landboukundige dimensies van droogte verteenwoordig.

Die model, geïmplementeer vir die periode 1981–2019 in die Suidwes-Kaap, het latente droogtetoestande geskat met behulp van die verwagting–maksimering (EM)-algoritme en die knoopboom (JT)-afleiding. Modelseleksie, gelei deur standaard kriteria, het 'n ses-toestand-konfigurasie geïdentifiseer. Uitsette is onttrek deur die Viterbi-algoritme en die Maksimum Posterior Marginale (MPM)-reël.

Resultate toon dat die DNBC groot historiese droogtes akkuraat geïdentifiseer het en die

hoogste F1-telling tussen insette behaal het, wat betroubare klassifikasie en gebalanseerde sensitiwiteit aandui, en sodoende die gebruik daarvan vir waarskynlikheidsgebaseerde droogtemonitering ondersteun.

Contents

List of Figures

List of Tables

Nomenclature

Variables and functions

T	Total number of time steps
t	A single time step index
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
π_i	Prior probability of the first latent drought random variable: $p(S_1 = i)$
$\boldsymbol{\theta}_1$	Set of all initial state probabilities
$a_{i,j}$	Transition probability from latent state i to j: $p(S_t = j \mid S_{t-1} = i)$
$\boldsymbol{\theta}_2$	Set of all transition probabilities
$b_i^{(n)}(j)$	Emission probability: $p(A_t^{(n)} = j \mid S_t = i)$
$\boldsymbol{\theta}_3$	Set of all emission probabilities
Θ	The complete set of all model parameters
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: $\{A_t^{(1)}, A_t^{(3)}, A_t^{(3)}\}$
$\mathbf{A}_{1:T}$	Sequence of all input variables across all time steps: $(\mathbf{A}_1, \ldots, \mathbf{A}_T)$
$\mathbf{S}_{1:T}$	Sequence 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 set of random variables $\mathbf{X}$ in the clique tree
$\delta_{i \to j}(\mathbf{S})$	Message from cluster i to j over separator set of random variables $\mathbf{S}$ in the clique tree
$p(X)$	Probability mass function of the discrete random variable X

Acronyms and abbreviations

PGM	Probabilistic graphical model
RV	Random variable
EM	Expectation–maximisation
MPM	Maximum posterior marginal
DNBC	Dynamic naive Bayes classifier
HMM	Hidden Markov model
JT	Junction tree
AIC	Akaike information criterion
BIC	Bayesian information criterion
SPI	Standard precipitation index
SDI	Streamflow drought index
NDVI	Normalised difference vegetation index
DWS	Department of Water and Sanitation
SAWS	South African Weather Service
NOAA	National Oceanic and Atmospheric Administration
UCT	University of Cape Town

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 the Western Cape severe droughts have left lasting socio-economic scars with notable events including 1973–74, 1983–84, 1991–92, 1994–95, 2000–2001, 2003–2004, 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 standardised 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 and reservoir storage. This type of drought typically lags behind meteorological drought and is measured using indices such as the streamflow drought index (SDI) or the standardised streamflow index (SSI), which are metrics derived from river monitoring [19–22].

Agricultural drought describes the phenomenon where the climate causes a significant decline in crop yield 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 normalised 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. Additionally, they have experimented with 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 climate, the East Coast sees summer rainfall and a subtropical climate, 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 creates conflicting information from each index, which complicates interpretation and leads to poor decisions. 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 drought, SDI for hydrological drought, and NDVI for agricultural drought. Relying on any single index therefore provides an incomplete view. These indices frequently contradict each other requiring industry experts to analyse them, ultimately leading to false positives and negatives for different users. This 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 United States drought monitor (USDM). 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].

1.1.5. Probabilistic Graphical Models and Their Value in Drought Monitoring

Probabilistic graphical models (PGMs) constitute a family of statistical models that combine principles from probability theory and graph theory to represent complex systems of interdependent variables. These models enable principled learning and inference under uncertainty by encoding joint probability distributions in a graphical form [38].

PGMs are powerful tools for creating composite drought indicators as they are able to handle the field's inherent uncertainties which include data sparsity, measurement errors, and complex climate relationships. Other notable approaches typically fall into two categories. The first involves *weighted aggregation*, where normalised drought indices are combined using fixed or subjectively assigned weights [39]. This method is simple, assumes independence between indices, and imposes a static linear relationship. This does a poor job at capturing the complexities of drought. The second approach employs *dimensionality reduction techniques*, with principal component analysis (PCA) being the most widely used [40]. PCA assumes linear relationships between variables, is sensitive to both outliers and scaling, and has poor interpretability. PGMs on the other hand offers a principled and an interpretable means of modelling both the dependencies and uncertainties of climatic conditions. When applied to the construction of composite indicators, PGMs can explicitly encode how different aspects of drought influence one another over time. Models like the dynamic naive Bayes classifier (DNBC), a variant of the hidden Markov model (HMM), extend this capacity by incorporating temporal dependencies between states. This enables the model to infer drought evolution using sequential data. Additionally, PGMs produce probabilistic outputs that quantify the likelihood or confidence of different drought classifications—something that can not be done using the other two methods mentioned.

Recent studies have been successfully incorporating DNBCs 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 [41, 42].

1.2. Problem Statement

South Africa lacks a composite drought indicator that integrates the meteorological, hydrological, and agricultural dimensions of drought. Existing systems tend to focus on individual indices which capture isolated aspects of drought but fail to represent the full complexity of the problem. As a result, decision-makers lack a cohesive view of drought and thus struggle to implement timely and effective intervention strategies.

The problem addressed in this study is therefore the absence of a probabilistically principled framework capable of combining heterogeneous drought indicators into a single, adaptive, and interpretable composite index. The desired outcome is a model that integrates three indices, that being the SPI, SDI, and NDVI, into a probabilistic structure capable of inferring drought state transitions over time.

The scope of this study is deliberately restricted to the meteorological, hydrological, and agricultural domains of drought. While the socio-economic dimension is important, it is excluded. This exclusion is primarily due to the considerable complexity involved in quantifying socio-economic indicators in the South African context. For existing methods, the limited availability of reliable, open-access data makes it near impossible. The model is applied to the southwestern Cape region of the country using openly available datasets covering the period 1981–2019.

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.

1.3. Overview of Project Design

This section provides a high-level overview of the design and methodology underpinning this project. The primary goal is to develop a probabilistic drought monitoring framework for South Africa based on a DNBC. The following subsections outline the complete workflow—from data acquisition and index derivation to model construction, state interpretation, and performance evaluation. The intention here is not to provide full theoretical or mathematical detail, but to clarify the logical sequence through which the final drought classification system was built and assessed.

The study spans the period 1981–2019 and focuses on areas in the southwestern Cape region (locations of data sources are plotted in Figure 3.1). All datasets used are open-source. Precipitation and river streamflow observations are collected, cleaned, and processed to compute the SPI and the SDI respectively. The SPI quantifies the short-term rainfall anomalies captured at weather stations while the SDI measures streamflow deficits at river gauging stations. Both of these are commonly used as statistical measures to capture the meteorological and hydrological aspects of drought respectively. In contrast, NDVI captures vegetation conditions, which serves as a proxy for agricultural drought. This data is obtained directly from the National Oceanic and Atmospheric Administration (NOAA) who maintain a global data record of NDVI across the globe [43]. These three indices were discretised and formatted to form the input for the DNBC.

The DNBC represents the evolution of drought conditions by linking the input indices (SPI, SDI, and NDVI) to a hidden or latent discrete RV at each time step. This structure allows the latent RVs to capture the true, underlying drought dynamics by integrating the different dimensions of each input index while also accounting for temporal dependencies.

Model training is performed using the expectation-maximisation (EM) algorithm, which iteratively estimates the model parameters that maximise the likelihood of the observed data (i.e., the input indices). Inference over the probabilistic structure is implemented via

the junction tree (JT) algorithm for efficient computation of joint probabilities.

The number of hidden drought states (i.e., the cardinality of the latent RV) is not fixed but rather determined through model selection, alongside the time scales (rolling window size) of both the SPI and SDI. Models with these varying hyperparameters are evaluated using the Akaike information criterion (AIC), Bayesian information criterion (BIC), and the maximised log-likelihood.

Once the optimal hyperparameters are identified, the model is then re-trained with this final structure. This yields, for each time step, a probability distribution over a finite set of discrete drought states. These states can be interpreted along a conceptual scale from dry to wet conditions, where the binned values below the midpoint represent progressively drier states and those above it progressively wetter states. Thus, rather than assigning a single categorical label, the model quantifies the severity of drought at each time step. Figure 1.1 shows a diagram illustrating how the input and output components work with the DNBC model where the number of latent drought states is 6 as an example.

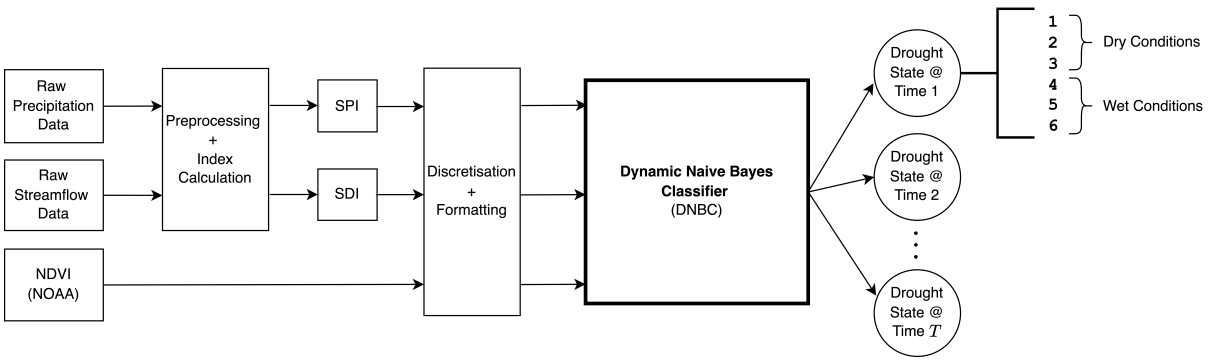


Figure 1.1: Schematic of the end-to-end project pipeline, showing data preprocessing, index calculation, DNBC input formatting, and final classification into six latent drought states as an example.

In order to extract the categorical drought classifications, the DNBC is decoded using two complementary approaches. The first is the Viterbi algorithm, which identifies the single most probable sequence of hidden drought states over the entire period, providing a coherent narrative of the drought's evolution. The second is the maximum posterior marginal (MPM) rule, which determines the most likely drought state for each individual time step, along with an associated probability that serves as a measure of uncertainty for the classification. Using both methods provides a more robust and nuanced classification of drought while being interpretable.

Finally, model performance is evaluated against the individual input indices (SPI, SDI, and NDVI) to assess whether the DNBC provides a more robust and cohesive representation of drought. Evaluation focuses on the model's ability to correctly identify major historical drought events in the study area (1983–1984, 1991–1992, 1994–1995, 2000–2001, 2003–2004, 2014–2016, and 2017–2018). Both quantitative and qualitative analyses are conducted. Temporal plots are used to visually compare drought classifications across indices, while

statistical metrics—including recall, precision, F1-score, and overall accuracy—are applied to objectively quantify performance.

This overview summarises the end-to-end design of the project, from data source to probabilistic classification. Later chapters will provide more clarity and detail regarding the theory, implementation, and results. For a visual representation of this design, see Appendix C.

1.4. Contributions

This project contributes to drought monitoring research in three distinct ways.

Firstly, it introduces a probabilistic, data-driven framework for composite drought monitoring in South Africa. This offers an alternative to other conventional methods which often lack interpretability. By integrating SPI, SDI, and NDVI the study demonstrates that PGMs can effectively capture cross-domain and temporal dependencies in drought evolution.

Secondly, it evaluates the DNBC's drought monitoring capability by comparing its detection of major historical droughts against trusted industry indices. This provides practical evidence for the model's operational feasibility.

Finally, this project contributes a framework for composite drought indicator development in South Africa using only open-source data while keeping reproducibility in mind. This work sets the stage for future early-warning systems that can adapt to uncertainty and provides a clear method for adding other indicators, like socio-economic data, later on.

In doing so, the study not only fills a critical gap in South African drought monitoring but also provides methodologies for composite indicator development in data-sparse environments.

1.5. Report Outline

This report is organised into four main sections, each addressing a key stage in the development and evaluation of the proposed composite drought indicator.

Chapter 2: Literature Review and Theoretical Framework

This chapter reviews existing approaches to composite drought monitoring and establishes the theoretical foundation for the modelling framework used in this study. The first part surveys prior research on multi-index integration, drought index evaluation for South African conditions, and the application of dimensionality-reduction techniques. The second part introduces PGMs, describing their structure, inference mechanisms, and learning algorithms.

Chapter 3: Methodology

Chapter 3 details the full project design pipeline, from data acquisition to model imple-

mentation. It begins with a description of the data sources, preprocessing steps, and the computation of the individual drought indices. This is followed by a detailed exposition of the DNBC model design, including its structure, inference procedures, parameter estimation, and model selection criteria. The chapter concludes by outlining the computational environment and the implementation of the end-to-end data and modelling workflow.

Chapter 4: Results

This chapter presents the outcomes of model training, evaluation, and analysis. It first discusses model selection and the definition of drought states, followed by an examination of model behaviour through latent-state sequences and confidence analysis. Quantitative evaluation metrics, including accuracy, precision, recall, and F1-score, are then used to assess model performance relative to the individual drought indices.

Chapter 5: Summary and Conclusion

The final chapter summarises the main findings of the study and evaluates the extent to which the research objectives were achieved. It reflects on the methodological and practical contributions of the DNBC-based framework to drought monitoring and concludes by identifying limitations and proposing directions for future research.

Chapter 2

Literature Review and Theoretical Framework

Advancing drought monitoring requires an understanding of both the prior research and theoretical foundation. This chapter addresses this need by first reviewing key studies that have shaped the development of composite drought indicators (Section 2.1). It then establishes the core theoretical framework of probabilistic graphical models (PGMs) that underpins this study (Section 2.2).

Together, these two perspectives—empirical and theoretical—motivate the use of PGMs as a promising tool for drought monitoring, while establishing the theoretical background for the rest of this report.

2.1. Literature Review on Drought Monitoring

Scholars have investigated methods ranging from traditional single-index approaches to more sophisticated probabilistic models. The following reviews four key studies that collectively sketch the evolution of composite drought indicator development. Two that applied dynamic naive Bayes classifiers (DNBCs) to integrate multiple drought indicators in South Korea; one that assessed the suitability of various drought indices for South Africa's complex climatic conditions; and one that employed a dimensionality reduction technique to construct a composite drought index for New Mexico.

The review summarises the objectives, methods, and findings of each study, followed by a synthesis that identifies the limitations of current approaches and motivates the use of PGMs for this research.

2.1.1. Integrating Multi-Index Drought Data Using a DNBC

This study by Kim et al. developed a dynamic naive Bayes classifier multiple drought index (DNBC-MDI) to generate a probabilistic and multi-dimensional assessment of drought risk. The core idea was to create one coherent model that would combine established drought indicators in order to capture the four most important dimensions of drought—that is, the standardised precipitation index (SPI) for meteorological drought, the streamflow

drought index (SDI) for hydrological drought, the evapotranspiration stress index (ESI) for agricultural drought, and the water supply condition index (WSCI) for socio-economic drought.

The model was applied to the Han River basin, using observed records from 1974–2016 and future climate projections extending to 2099 under a high-emission scenario. The DNBC was trained to recognise hidden drought states over time, adjusting its internal parameters through an iterative learning process known as expectation-maximisation (EM). This allowed the model to infer the likelihood of different drought conditions given observed changes in precipitation, streamflow, and evapotranspiration.

To assess the relationship between different drought variables and their joint risk, the authors used a statistical function known as the *Clayton copula*. This is a mathematical way of linking multiple probability distributions so that their dependencies can be captured and analysed together, rather than assuming each drought indicator behaves independently. Using this approach, they estimated long-term drought risks such as 100-year return periods, representing the probability of extreme drought events over time.

The results showed that the DNBC-MDI achieved higher classification accuracy than any of the individual indices on their own, while successfully reproducing several major historical drought events (1994–1995, 2001, 2008–2009, 2012, 2014–2015). The authors also highlighted two main limitations: first, the model relied mainly on climate simulations and did not incorporate remote-sensing data such as satellite-derived vegetation indices; and second, it assumed that the input drought indices were conditionally independent of one another. This is an assumption that may not fully hold in reality given the physical links between rainfall, river flow, and evaporation.

Overall, this paper demonstrated the feasibility and potential of DNBC-based models for integrating multiple drought indicators into a single probabilistic framework, while also highlighting the importance of addressing data diversity and inter-variable dependencies when adapting such methods to new contexts [41].

2.1.2. Probabilistic Frameworks for Drought Characterisation with DNBCs

Similarly to the study above, Chen et al. also applied a DNBC to create a composite drought indicator. However, in this paper the researchers only utilised three indicators which were SPI, SDI and the normalised vegetation supply water index (NVSWI) which attempts to capture agricultural drought and is thus an alternative to the ESI. The aim was to evaluate whether this model could capture drought events more accurately and consistently than any single index on its own, particularly in terms of detection, classification, and persistence through time. It is important to note the omission of the socio-economic aspect of drought here.

The model was applied to the upper Han River Basin in South Korea using observations from 1980–2015 and satellite-based data from 2003–2015. Within this framework, the DNBC represented drought as a sequence of hidden states, each corresponding to a distinct drought severity level. These hidden states evolved over time in response to changes in the observed indices. The number of drought states was chosen using statistical model selection criteria to balance model complexity and interpretability. Parameter estimation relied on the EM algorithm, similar to the study above.

The results showed that the DNBC successfully reproduced several well-documented drought episodes (2004, 2006, 2008–2009, 2014, 2015) and accurately reflected both drought duration and persistence. In comparative testing, the DNBC-based drought states detected nearly all events identified by the individual indices, matching or exceeding their detection rates across the different drought types. The analysis also revealed that meteorological and hydrological indicators were more closely related to one another than to the agricultural indicator, illustrating the complexity of cross-sector drought linkages. Overall, the DNBC provided a coherent probabilistic framework that incorporated uncertainty directly into drought monitoring.

Nonetheless, the authors noted several limitations. The model was limited to three indices, omitting potentially informative climatic variables such as temperature, water vapour, and solar radiation, which may have strengthened its predictive capacity. Furthermore, the DNBC assumed conditional independence among its input indicators, an assumption that may oversimplify the real-world interconnections between atmospheric, hydrological, and vegetation processes.

Despite these challenges, the study clearly demonstrates how DNBCs can unify multiple drought indicators and provides a methodological framework for developing composite indices, thus forming the foundation of this study [42].

2.1.3. Evaluating Drought Indices for South African Conditions

Mukhawana et al. conducted a review that evaluated existing drought indices in order to identify the most effective and feasible ones for integrated drought monitoring in South Africa's diverse climate. It examined eight widely used indicators, each representing different aspects of drought. Their goal was to determine which indicators could realistically be applied in the South African context, given the country's sparsity of data.

Following the World Meteorological Organisation's (WMO) 2016 guidelines, the review assessed each index against five criteria: data requirements, computational simplicity, sensitivity to drought conditions, adaptability for integration, and overall reliability. The evaluation was based on existing studies conducted across South Africa and other regions with comparable climates.

The findings revealed that indices relying on detailed soil and surface-water data (such

as the Palmer drought severity index and surface water supply index) are not feasible due to the lack of consistent data. In contrast, indicators derived from rainfall, and satellite observations (such as the standardised precipitation index, standardised precipitation-evapotranspiration index, vegetation condition index, and related streamflow measures) were found to be both practical and sufficiently sensitive to regional drought patterns. However, the review highlighted technical challenges, such as computational difficulties arising from missing values in input data, as well as the absence of reliable groundwater records that would be used for validation.

The study emphasised that no single index can fully represent all drought dimensions, advocating for a multivariate approach that combines complementary indicators. Additionally, it directly informed this project's selection of data-efficient, complementary, and reproducible drought indices as inputs for the DNBC model [37].

2.1.4. Principal Component Analysis for Composite Drought Indicators

The following study, by Poudel et al., developed a composite drought indicator in the context of New Mexico (CDI-NM) using principal component analysis (PCA) to integrate multiple satellite-derived variables, including rainfall, land surface temperature, and vegetation indices. The aim was to construct a data-driven tool capable of identifying historical drought events and assessing drought extent across the state.

Using datasets from 2003–2019, the authors applied PCA on a monthly basis and validated the suitability of the extracted components using standard checks for sampling adequacy and variable interdependence (using the Kaiser-Meyer-Olkin and Bartlett tests). Model performance was evaluated against the SPI and agricultural yields. The CDI-NM correlated strongly with both these measures, effectively capturing known drought periods between 2003 and 2018.

However, the study's methodology exposes several recurring weaknesses of PCA-based composite indicators. The technique assumes linearity and temporal stationarity, which constrains its capacity to capture dynamic, non-linear dependencies between input variables. Its reliance on fixed weighting structures further limits adaptability to evolving drought conditions. Additionally, redundancy among vegetation indices and sensitivity to scaling introduce potential distortions in the resulting indicator. These limitations highlight a core shortcoming of dimensionality reduction methods—they capture covariation, not causal relationships.

This gap directly motivates the use of PGMs, as they provide a more interpretable framework capable of capturing the causal interdependencies and temporal dynamics among input variables using DNBCs [40].

2.1.5. Conclusion

The reviewed literature illustrates the growing movement toward integrated, probabilistic approaches for drought monitoring. The first two studies established the methodological foundation for DNBC models, demonstrating their capacity to integrate multiple drought indicators while handling uncertainty and variable interdependencies. The third study provided crucial context with regards to which indices are feasible to obtain in the South African context. Finally, the fourth study revealed the limitations of conventional statistical integration methods such as PCA, which rely on linear and static relationships, and highlighted the need for models that can explicitly represent conditional dependencies and temporal dynamics.

These studies and reviews show a clear trajectory from static, correlation-based methods toward dynamic, probabilistic frameworks for drought monitoring. 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.

2.2. Probabilistic Graphical Model Background

Probabilistic graphical models (PGMs) provide a framework that sits at the intersection of probability theory and graph theory. In a PGM, a graph structure compactly encodes a complex probability distribution, where nodes represent random variables (RVs) and edges correspond to direct probabilistic interactions between them. This paradigm enables the effective construction and application of models for complex probabilistic reasoning [38, 44].

2.2.1. Factors and Bayesian Networks

One way to view a graphical model is as a structured factorisation of a large joint probability distribution. Rather than representing the probability of every possible assignment to all variables in the domain, the model "breaks up" this high-dimensional space into a collection of smaller components—these are called *factors*. A factor is a function defined over one or more RVs that assigns a numerical score to each possible combination of their values. This score reflects how likely that combination is and is thus analogous to a standard probability distribution. However, the scores of a factor need not sum to one, when they do they are considered normalised and become proper probability distributions. By combining a set of locally defined factors according to the graph structure, the global joint distribution over all variables can be reconstructed while maintaining a compact and interpretable representation.

A common and powerful class of PGMs is the *Bayesian network (BN)*, where the graph is directed and acyclic. Each directed edge represents a conditional dependence of a

child node given its parent nodes. The network can be viewed as a collection of factors, where each factor corresponds to the conditional probability distribution of a variable given its parents. For example, in a simple three-node network: $A \rightarrow B \rightarrow C$, the joint probability distribution is constructed by multiplying all the factors in the graph as seen in Equation (2.1).

$$p(A, B, C) = p(A)\, p(B \mid A)\, p(C \mid B) \tag{2.1}$$

This decomposition shows how the overall model is built from locally defined factors, yielding a compact and interpretable representation. Once this joint distribution is specified, any probabilistic query about the system can be answered through inference, as discussed next.

Once the joint distribution is defined, *inference* can be perfomed—that is, determine the probability of any RV or set of RVs given the observations of others. This can be achieved through marginalisation, where irrelevant RVs are summed out. Using the previous example, computing $p(C \mid A)$ is achieved by marginalising over B and applying Bayes' rule as shown below.

$$\begin{aligned} \frac{\sum_B p(A, B, C)}{p(A)} &= \frac{p(A, C)}{p(A)} \\ &= p(C \mid A) \end{aligned}$$

This process allows exact probabilistic reasoning without exhaustively enumerating all combinations of RV values.

2.2.2. Efficient Inference with Junction Tree Algorithm

For models of even moderate size, directly computing the joint distribution by multiplying all factors becomes intractable—that is, computationally too complex to solve in a reasonable amount of time. In such cases, exact inference can be achieved using the *junction tree (JT)* algorithm.

This method reorganises the BN into a specific data structure called a junction tree, sometimes called a clique tree, where each node represents a cluster, or clique, of RVs ($\psi_i(\mathbf{A})$). Any two clusters ($\psi_i(\mathbf{A})$ and $\psi_j(\mathbf{B})$) are connected such that the scope of each edge, known as a sepset, consists of the RVs shared between the two adjacent clusters ($\mu_{ij}(\mathbf{A} \cap \mathbf{B})$).

Any valid JT must satisfy two key properties: the *family preservation property*, which ensures that each factor in the original network is assigned to a cluster that contains its scope, and the *running intersection property*, which guarantees that for any RV present in two clusters, all clusters along the unique path between them must also contain that RV [38].

The inference process involves a coordinated scheme of message-passing, where adjacent

clusters in the tree communicate with each other. This communication is achieved by passing *messages*—functions that encapsulate the relevant probabilistic information or, more conceptually, the "belief" from one cluster to another. The purpose of this is to allow each cluster to update its local beliefs in a way that is globally consistent across the whole tree.

Messages are passed according to a specific schedule which begins at the leaf clusters of the tree and sends messages inward. A cluster will only pass a message to its neighbour once it has received messages from all its other neighbours. This ensures that information is propagated correctly, ultimately leading to a state of global consistency known as *calibration*. Once the tree is calibrated, the updated cluster potentials and sepsets represent the true marginal posterior distributions, enabling exact, tractable inference [38, 44, 45].

2.2.3. Parameter Estimation with Expectation-Maximisation

Finally, when the parameters of the model, such as the distributions governing its factors, are unknown, parameter estimation is required. The overarching goal of parameter estimation is typically to maximise the data log-likelihood and several approaches exist for this purpose, including maximum likelihood estimation (MLE), Bayesian inference, variational methods, or gradient-based optimisation techniques. The choice of approach depends on factors such as the structure of the model, the presence of latent RVs, and computational constraints.

When models involve latent RVs (these are RVs that are never observed) or incomplete data, direct optimisation of the likelihood becomes intractable. Although gradient-based methods are applicable, they are computationally expensive. The expectation-maximisation (EM) algorithm is frequently preferred in the context of PGMs, as it is specifically designed for likelihood optimisation with latent RVs. EM provides a stable, iterative framework that yields improvements at each step. The algorithm proceeds by alternating between two core steps:

(i) The expectation or E-step, which computes the expected values of the latent variables given the observed data and current parameter estimates.

(ii) The maximisation or M-step, treats the calculated expected sufficient statistics as observed, and performs maximum likelihood estimation to derive a new set of parameters.

This process repeats until convergence, yielding parameter estimates consistent with both the observed data and the probabilistic structure of the model [38, 46].

Together, these concepts form the theoretical basis for the DNBC proposed in this study, which combines temporal dependence with probabilistic reasoning to model drought conditions.

Chapter 3

Methodology

This chapter describes the methodology developed to build and apply the dynamic naive Bayes classifier (DNBC) for composite drought monitoring in the the southwestern Cape region. The primary objective was to create a model that integrates meteorological, hydrological, and agricultural drought indices into a unified, interpretable measure of drought conditions.

The process began with data acquisition and preprocessing, where the necessary input variables were sourced, cleaned, and transformed into discrete drought indices. Subsequently, the core DNBC model was developed, specifying its probabilistic structure and the procedures for inference and parameter learning. Finally, the model was implemented using appropriate computational tools and algorithms.

3.1. Data and Index Construction

3.1.1. Overview of Input Indices

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 first being the *standardised precipitation index (SPI)* which measures how much precipitation deviates from the long-term average making it a useful statistical measure for meteorological drought [16]. The second chosen index is the *streamflow drought index (SDI)*. Similar to the SPI, it measures how much river streamflow deviates from long-term averages and thus helps quantify hydrological drought [20]. Finally, the *normalised difference vegetation index (NDVI)* is a remote-sensing indicator widely used to monitor vegetation health and stress. This serves as a proxy for agricultural drought [24]. 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, 41, 42]. For a summary of these input indices, see Table 3.1.

Table 3.1: Selected drought indices and their characteristics.

Name	Acronym	Drought Aspect	Primary Data
Standardised precipitation index	SPI	Meteorological	Precipitation
Streamflow drought index	SDI	Hydrological	River streamflow
Normalised difference vegetation index	NDVI	Agricultural	Remote-sensing

3.1.2. Data Sources

Monthly precipitation data was obtained from an open source dataset authored by Conradie et al. and maintained by the University of Cape Town (UCT) [47]. This dataset provides rainfall values captured at weather stations across the southwestern region of South Africa and spans the period 1979–2019 in CSV format. This data offers consistency and a high degree of granularity.

Streamflow records were sourced from a collection of river gauging stations managed by the Department of Water and Sanitation (DWS), which maintains audited historical hydrology data [48]. The dataset provides measurements from individual locations across the country, with records spanning periods anywhere between 1903 and 2025. The daily data, accessible only in a text format on the DWS website, was scraped and converted to a structured CSV format.

NDVI was not computed but rather obtained directly from the "NOAA Climate Data Record (CDR) of AVHRR Normalised Difference Vegetation Index (NDVI), Version 5" dataset [43]. This dataset offers daily records that span the period 1981–2025, covers the entire globe in rasterised grids, and is provided in NetCDF format.

With the three primary data sources (precipitation, streamflow, and satellite-derived vegetation indices) acquired and converted into consistent, usable formats, the foundational data preparation is complete. The subsequent phase involves preprocessing this dataset for input into the dynamic naive Bayes classifier (DNBC) model.

3.1.3. Preprocessing and Cleaning

The first step in preprocessing was to establish the study period as well as the temporal resolution. The study was constrained to the overlapping period of the three data sources, which spans 1981–2019. Since the DNBC works on monthly time steps, all the data was processed to a monthly resolution to ensure temporal consistency across indices.

The preprocessing of the SPI and SDI data sources was largely similar, with a few key distinctions. Both rainfall and streamflow data sources were subjected to the following station filtering rules:

- Stations that did not span the full 1981–2019 study period were removed.
- Stations containing any missing values within this period were removed.
- Outlier detection was applied using the interquartile range (IQR) rule. This meant removing any stations that had a record which deviated by more than $1.5 \times IQR$ from the median of surrounding stations for that time step.

Although this criterion seems strict, the high density of available stations allowed for 205 weather stations and 154 river gauging stations to be left remaining. These clean stations collectively define the study area and are shown in Figure 3.1. Finally, the daily streamflow data was aggregated to monthly time steps, aligning with the temporal resolution of the raw precipitation data.

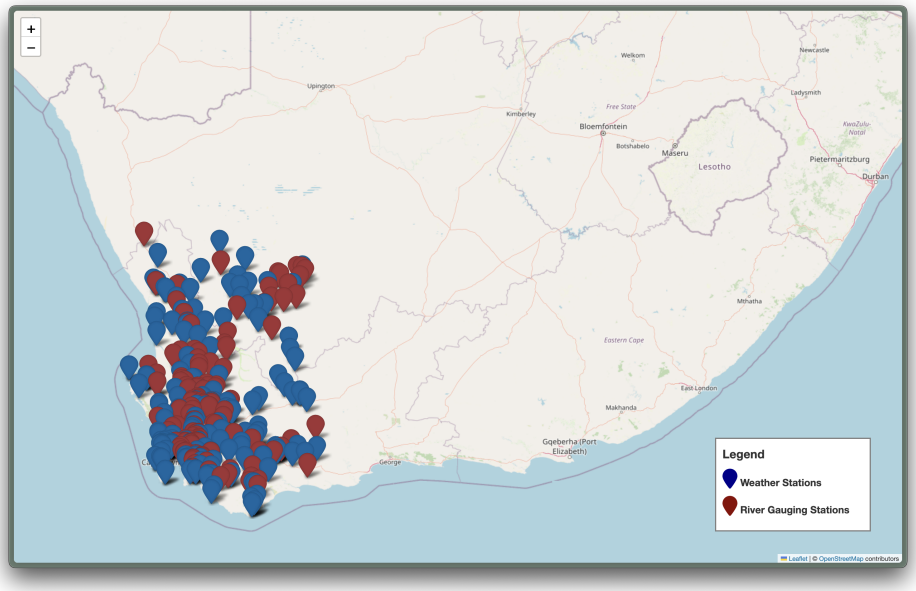


Figure 3.1: Locations of weather (blue) and river gauging (red) stations used in this study.

The preprocessing workflow for the NDVI data was comparatively simple. It required subsetting the global dataset to the established study area and defined study period, recalibrating the scaled integers to the standard floating-point range (-1.0 to 1.0), and aggregating the daily data to monthly averages. The results were then exported from NetCDF to CSV to align with the data format of the other indices.

3.1.4. Index Computation

With all datasets now cleaned, temporally aligned, and aggregated to monthly resolution, the next step involved computing the drought indices. No further computation was required for NDVI following the preprocessing stage. This subsection therefore focuses on the computation of the SPI and SDI.

Theoretical Background

Both SPI and SDI are statistical measures that transform raw observations into standardised indicators of drought intensity [16,20]. The procedure for each index is conceptually similar and can be summarised as follows:

1. Window-Based Aggregation: Observations are first aggregated over a specified temporal window to capture drought persistence at different timescales. Common practice varies between 3-, 6-, 9-, or 12-month windows, each reflecting short to long term drought characteristics.

2. Distribution Fitting: For each station and calendar month, the aggregated series is assumed to follow a parametric probability distribution. This accounts for the seasonality present in both rainfall and streamflow data. The fitted distribution is then used to compute the cumulative probability of each observed value, which expresses the rarity of that observation relative to values from the same calendar month in other years. In other words, it quantifies how far current conditions deviate from the historical norm for that time of year. Various distributions has been proposed in the literature, including gamma, Weibull, and log-normal [37].

3. Standard Normal Transformation: Once the cumulative probability $p(x)$ of an observation is obtained, it is transformed into a standard normal distribution (Z-score). This produces continuous, standardised values with mean zero and unit variance. Negative values indicate conditions that are drier than usual, while positive values represent wetter than usual. The resulting Z-scores form the SPI or SDI, depending on the underlying input data (precipitation or streamflow).

Implementation in This Study

In this study, both the SPI and SDI were computed independently for each station using the gamma distribution, keeping consistent with standard practice in drought monitoring literature [37]. While most studies employ fixed aggregation windows, this project treated the aggregation window length as a hyperparameter that was determined during model selection.

3.1.5. Discretisation of Indices

Once again, all three input indices—SPI, SDI, and NDVI—are all continuous measures. However, the proposed model requires discrete inputs. Accordingly, each index was discretised into categorical bins based on thresholds widely used in literature. This discretisation not only aids in model implementation but it also motivates interpretability. Table 3.2 below illustrates the bins used.

Table 3.2: 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	$0.6 \leq x < 1$

At this stage, all three drought indices were computed, discretised, and temporally aligned over the 1981–2019 period. These indices together formed the input set for the DNBC.

3.2. Model Development

This section details the development of the DNBC for this project. The model development process proceeds through five key stages: design, where the model is fully specified; inference, where algorithms for answering queries about the model are formalised; parameter estimation, where model parameters are learned from data; model selection, where various hyperparameter configurations are evaluated to identify the most representative model; and finally, output decoding, where the model results are exported in two forms. It is important to note that this section does not introduce new theoretical contributions but rather directly implements established DNBC theory as found in existing literature [41, 42].

3.2.1. Model Design

Defining the Random Variables

The proposed DNBC is constructed with 3 input random variables (RVs) observed across T discrete time steps. All 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 integer values of S_t are ordinally scaled, representing a spectrum of drought severity from severe drought (lower values) to wet conditions (higher values), with the central values typically indicating a neutral or normal state.

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

$$A_t^{(n)} \in \{1, 2, \ldots, 5\}, \quad n = 1, 2, 3, \quad t = 1, \ldots, T.$$

The meanings attached to each value follows directly from the discretisation shown above in Table 3.2. For clarity, below is the assignment of these input indices which constitute the dataset $\mathcal{D}$.

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

We define the following notation which will be used throughout the model formulation:

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

where each $\mathbf{A}_t = \{A_t^{(1)}, A_t^{(2)}, A_t^{(3)}\}$.

Graphical Structure and Joint Distribution

Figure 3.2 below displays the model diagram for the DNBC model as a Bayesian network (BN) with T time steps and 3 input RVs at each time step.

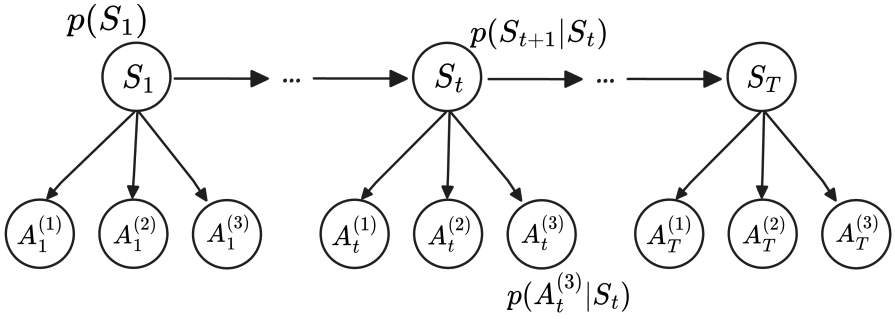


Figure 3.2: The DNBC can be represented as a Bayesian network (BN) unfolding over time. At each time step t, a latent drought state S_t is modelled as a discrete random variable that governs the latent structure, while the observed input variables $\mathbf{A}_t = \{A_t^{(1)}, A_t^{(2)}, A_t^{(3)}\}$ are each solely dependent on S_t

As discussed in Section 2.2, we can utilise this graph to construct the joint probability distribution which is shown in Equation (3.1) below.

$$p(\mathbf{S}_{1:T}, \mathbf{A}_{1:T}) = p(S_1) \cdot \prod_{t=1}^{T-1} p(S_{t+1} \mid S_t) \cdot \prod_{n=1}^{3} \prod_{t=1}^{T} p(A_t^{(n)} \mid S_t) \tag{3.1}$$

Parameterising the Model

Following established literature, the DNBC is fully specified by three sets of parameters, namely the prior, transition, and emission probabilities.

Prior Probabilities: The prior probabilities, denoted as $\boldsymbol{\theta}_1$, represents the initial belief over the latent drought states before any observations are made. This distribution captures the likelihood of the system starting in each possible state at time $t = 1$. Formally, the prior probability for state i is defined as:

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

where $\sum_{i=1}^{m} \pi_i = 1$. These parameters are captured in the factor table shown in Table 3.3.

Table 3.3: Factor table for the prior probabilities ($\boldsymbol{\theta}_1$)

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

Transition Probabilities: Denoted as $\boldsymbol{\theta}_2$, this set of parameters define the likelihood of the system moving from one latent drought state to another between consecutive time steps. Formally, the probability of transitioning from state i at time t to state j at time $t + 1$ is denoted as:

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

where $\sum_{j=1}^{m} a_{i,j} = 1$ for all i. These parameters determine the model's temporal dynamics and can be presented both as a factor table and a transition matrix, shown in Table 3.4.

Table 3.4: Factor table for the transition probabilities ($\boldsymbol{\theta}_2$) and transition matrix $\mathbf{P}^1$.

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	$a_{m,m}$

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

Emission Probabilities: The relationship between the latent drought states and the observed drought indices at each time step is defined by the emission probabilities, denoted as $\boldsymbol{\theta}_3$. These parameters quantify the likelihood of observing a specific value for a given input index (SPI, SDI, or NDVI) when the system is in a particular hidden drought state.

Formally, the probability of observing value j for the n-th input RV at time t, given the latent state i, is defined as:

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

This set of conditional probabilities is presented as a factor table, shown below in Table 3.5.

Table 3.5: Factor table for the emission probabilities ($\boldsymbol{\theta}_3$)

$A_t^{(n)}$	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	$b_2^{(n)}(2)$
⋮	⋮	⋮
5	m	$b_m^{(n)}(5)$

The full set of these parameters, $\Theta = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$, fully determine the complete DNBC. A key characteristic of this model is that the parameters are time-invariant, meaning the probability of transitioning between drought states and the likelihood of observing specific index values remain constant irrespective of the time step t. This is one of the model's core assumptions, which is discussed in the following section.

Assumptions

Finally, it is important to note the three assumptions the DNBC is built on, which defines its structure but also impose inherent limitations:

(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 the entire time series.

(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.

These assumptions have important practical implications. The first-order Markov assumption (i) implies that the model has a very short "memory", only using the information from the previous time step to determine the current one. Stationarity (ii) simplifies the

model by requiring only a single set of transition and emission probabilities for the entire period, as seen in the previous section. Most critically, the conditional independence assumption (iii) means that if the true, underlying drought state S_t is known, then knowing the value of one indicator (for example the SPI) provides no additional information about the value of another (such as the SDI or NDVI). In other words, the latent state S_t fully accounts for all the dependencies between the observed indices. Additionally, this third assumption allows for the following factorisation which will become useful at a later stage.

$$\begin{aligned} p(\mathbf{A}_t \mid S_t) &= p(A_t^{(1)}, A_t^{(2)}, A_t^{(3)} \mid S_t) \\ &= p(A_t^{(1)} \mid S_t) p(A_t^{(2)} \mid S_t)(A_t^{(3)} \mid S_t) \\ &= \prod_{n=1}^{3} p(A_t^{(n)} \mid S_t) \end{aligned} \tag{3.2}$$

3.2.2. Inference

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

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

This will later be used for the E-step in the expectation-maximisation (EM) algorithm.

The inference procedure is carried out using the junction tree (JT) algorithm which provides a means to obtain exact inference efficiently. The first step is to construct a JT that adheres to both the family preservation property as well as the running intersection property illustrated in Figure 3.3.

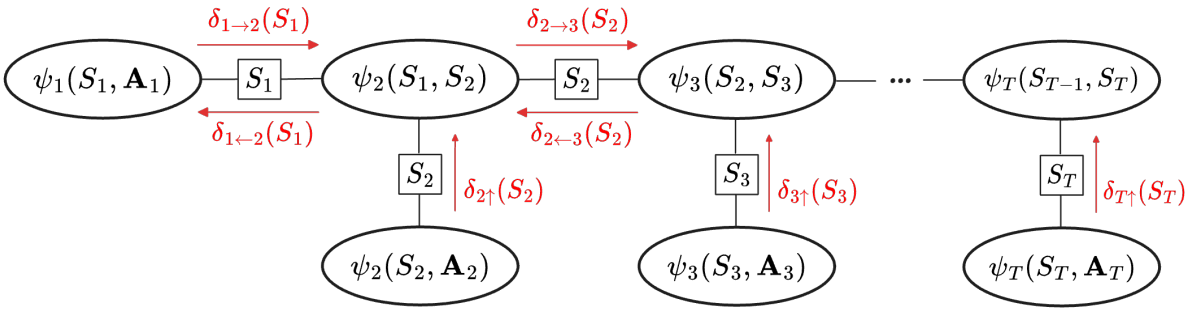


Figure 3.3: Junction tree representation of the DNBC. Each cluster groups together latent state variables and observed input indices, with sepsets defined along the edges. Messages are propagated through the tree to perform exact inference.

Each cluster in the tree represents a subset of RVs from the original DNBC. The relationship between these variables within a cluster is defined by its cluster potential, shown in Table 3.6. This structure relies on the factorisation of the observed indices given

in Equation (3.2).

Table 3.6: Cluster potentials for the junction tree of the proposed model.

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

Message-Passing

Once the JT is defined, the next step is message-passing. Recall from Section 2.2.2 that, conceptually, a message represents the "belief" of one cluster about the variables it shares with another. As illustrated in Figure 3.3, messages are passed directionally between clusters and are defined over the scope of their sepset.

For message-passing, either the belief update (BU) or belief propagation (BP) approach may be followed. While both are equivalent in terms of computation within the JT algorithm, BP was used because it is conceptually simpler.

Since the objective is to compute the posterior distribution $p(\mathbf{S}_{1:T} \mid \mathbf{A}_{1:T}, \Theta)$, only the clusters $\psi_t(S_t, S_{t+1})$ and their corresponding sepsets $\mu_{t,t+1}(S_t)$ need to be fully calibrated. Thus, downward messages—those from $\psi_t(S_{t-1}, S_t)$ to $\psi_t(S_t, \mathbf{A}_t)$—are not required for this task. Message definitions are therefore restricted to those shown in Figure 3.3, which are sufficient for obtaining the desired marginals.

Full theoretical derivations are omitted here for brevity, as the necessary conceptual foundation has already been provided. The following therefore presents only the defined messages used in this study.

Upward messages: Because the observed attribute variables $\mathbf{A}_t$ are no longer random, the marginalisation over their possible values becomes redundant. The upward messages therefore collapse to likelihood terms that express how compatible the observed evidence is with each latent state S_t. This yields the following simplified form:

$$\begin{aligned} \delta_{t\uparrow}(S_t) &= \sum_{\mathbf{A}_t} \psi_t(S_t, \mathbf{A}_t) \\ &= \sum_{\mathbf{A}_t} p(\mathbf{A}_t \mid S_t) \\ &= p(\mathbf{A}_t \mid S_t) \end{aligned} \tag{3.3}$$

Rightward messages: Once again, In the JT algorithm message propagation begins at the leaf clusters and proceeds inward. Accordingly, rightward messages originate at the leftmost cluster. The initial rightward message, Equation (3.4), reflects the joint contribution of the prior distribution on S_1 and its associated observation likelihood. Thereafter, messages are recursively propagated through time as shown in Equation (3.5). This process effectively integrates information from previous states and the current observation.

$$\begin{aligned}\delta_{1\to 2}(S_1) &= \sum_{\mathbf{A}_1} \psi_1(S_1, \mathbf{A}_1) \\ &= \sum_{\mathbf{A}_1} p(S_1)p(\mathbf{A}_1 \mid S_1) \\ &= p(S_1)p(\mathbf{A}_1 \mid S_1) \end{aligned} \tag{3.4}$$

$$\begin{aligned}\delta_{t\to t+1}(S_t) &= \sum_{S_{t-1}} \psi_t(S_{t-1}, S_t)\delta_{t-1\to t}(S_{t-1})\delta_{t\uparrow}(S_t) \\ &= \sum_{S_{t-1}} p(S_t \mid S_{t-1})\delta_{t-1\to t}(S_{t-1})p(\mathbf{A}_t \mid S_t) \\ &= p(\mathbf{A}_t \mid S_t) \sum_{S_{t-1}} p(S_t \mid S_{t-1})\delta_{t-1\to t}(S_{t-1}) \end{aligned} \tag{3.5}$$

Leftward messages: Similarly, leftward propagation begins at the final cluster and proceeds backward through time. The initial message from the last time step, Equation (3.6), combines the transition and emission probabilities at T, while the recursive form, Equation (3.7), propagates backward through time in the same way that rightward messages move forward.

$$\delta_{T-1\leftarrow T}(S_{T-1}) = \sum_{S_T} p(S_T \mid S_{T-1})p(\mathbf{A}_T \mid S_T), \tag{3.6}$$

$$\delta_{t-1\leftarrow t}(S_{t-1}) = \sum_{S_t} p(S_t \mid S_{t-1})\delta_{t\leftarrow t+1}(S_t)p(\mathbf{A}_t \mid S_t). \tag{3.7}$$

Together, These three sets of messages—upward, rightward and leftward—when passed in the correct ordering yield a fully calibrated JT. At calibration, the cluster potentials and sepsets represent the true marginal distributions: $\psi_t(S_{t-1}, S_t) = p(S_{t-1}, S_t)$ and $\mu_{t,t+1}(S_t) = p(S_t)$. This provides tractable computation of the target posterior distribution $p(\mathbf{S}_{1:T} \mid \mathbf{A}_{1:T}, \Theta)$.

Forward-Backward Equivalence

At this stage it is natural to note that the procedure described above is a generalisation of the classical *forward-backward* algorithm for hidden Markov models (HMMs), which dominate the literature [41, 42]. The inward and outward messages of the JT formulation

are algebraically equivalent to the forward and backward recursions respectively [49–51]. The distinction lies solely in formulation where the JT framework provides a general inference architecture for arbitrary graphical models and the forward-backward algorithm is the special case applied to chain-structured models such as HMMs and DNBCs.

As a result of this equivalence, the rightward messages of the JT correspond to the forward quantities $\alpha_t = p(\mathbf{A}_{1:t}, S_t)$, while the leftward messages correspond to the backward quantities $\beta_t = p(\mathbf{A}_{t+1:T} \mid S_t)$. When these messages are combined at a cluster or sepset, the resulting marginal distributions $p(S_t \mid \mathbf{A}_{1:T})$ and $p(S_t, S_{t+1} \mid \mathbf{A}_{1:T})$ are identical to those obtained through the forward-backward algorithm.

Connection to Baum-Welch: This parallel extends directly to the *Baum-Welch* algorithm [52], which is the EM implementation for parameter learning in HMMs. In Baum-Welch, the E-step is performed using the forward-backward quantities, and the M-step updates the transition and emission probabilities to maximise the expected log-likelihood. 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.

This link provides a smooth transition to the following subsection, where parameter estimation for the model is discussed using the EM framework paired with the JT algorithm.

3.2.3. Parameter Estimation

The next step involves estimating the parameters which is required to align the model with the observed data for subsequent inference. Because the DNBC includes latent RVs that are never observed, the expectation-maximisation (EM) algorithm was used. Again, the following subsection details the application of established literature and theory revolving PGMs [38, 46]. As discussed in Section 2.2.3, EM provides an efficient iterative procedure that alternates between computing the expected sufficient statistics of the latent states (E-step) and updating parameter estimates to maximise the data log-likelihood (M-step). This approach is particularly suited to models such as the DNBC, where the latent state sequence must be inferred jointly with the model parameters.

It is necessary, before detailing the EM steps, to establish the following notation which consolidates terms from previous sections. Firstly being the hidden states, secondly the observed RVs or the dataset of the problem, and lastly the full set of model parameters (the prior, transition, and emission probabilities). This notation is shown below.

$$\mathcal{H} = \mathbf{S}_{1:T} = (S_1, S_2, \ldots, S_T)$$
$$\mathcal{D} = \mathbf{A}_{1:T} = (\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_T)$$
$$\Theta = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$$

1. E-Step: In this step, the current parameter estimates Θ are held fixed while we compute the posterior distribution over the latent state sequence:

$$q(\mathcal{H}) = p(\mathcal{H} \mid \mathcal{D}, \Theta) = p(\mathbf{S}_{1:T} \mid \mathbf{A}_{1:T}, \Theta). \tag{3.8}$$

This corresponds directly to the inference problem discussed previously in Section 3.2.2. The discussed JT algorithm yields the required marginal and pairwise posteriors, namely $q(S_t)$ and $q(S_t, S_{t+1})$, without explicitly computing the full joint distribution in Equation (3.8). These quantities are sufficient for parameter re-estimation in the following M-step.

2. M-Step: In the M-step, the choice of distribution $q(\mathcal{H})$ stays fixed to what was determined in the E-step—meaning the posteriors $q(S_t)$ and $q(S_t, S_{t+1})$ remains constant. The parameters Θ are then updated by maximising the expected complete-data log-likelihood:

$$\mathcal{Q}(\Theta) = \sum_{\mathcal{H}} q(\mathcal{H}) \cdot \log\, p(\mathcal{D},\, \mathcal{H} \mid \Theta).$$

Expanding this expression reveals contributions from the prior, transition, and emission components of the model. A full derivation of these terms is not the focus here and can be found in standard explanations of the EM algorithm. The key idea to grasp is that the maximisation over Θ decomposes naturally across its components $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, and $\boldsymbol{\theta}_3$, leading to the standard re-estimation update rules [53, 54]:

$$\boxed{\pi_i^{\text{new}} = q(S_1 = i)} \tag{3.9}$$

$$\boxed{a_{i,j}^{\text{new}} = \frac{\sum_{t=1}^{T-1} q(S_t = i, S_{t+1} = j)}{\sum_{t=1}^{T-1} q(S_t = i)}} \tag{3.10}$$

$$\boxed{b_i^{(n)}(j)^{\text{new}} = \frac{\sum_{t=1}^{T} q(S_t = i) \cdot \mathbf{1}(A_t^{(n)} = j)}{\sum_{t=1}^{T} q(S_t = i)}} \tag{3.11}$$

Practical Considerations

The EM algorithm requires an initial distribution over the model parameters to begin the iterative process. Its performance is highly sensitive to these initial conditions, as the algorithm may converge to a local rather than a global optimum. Consequently, the choice of initialisation can substantially influence the quality of the final estimates. Initialisation strategies typically involve random initialisation according to either a uniform or Gaussian

distribution [46].

A termination criterion is necessary to prevent overfitting, since the algorithm guarantees improvement of the likelihood with each iteration, even when doing so yields negligible improvements. Behaviourally, the EM algorithm typically performs large updates in early iterations, followed by progressively smaller refinements as posterior distributions stabilise. While EM acts as a useful tool for parameter learning in the DNBC, it remains sensitive to poor initialisation and can also converge slowly.

3.2.4. Model Selection

Model selection is a critical step in ensuring that the DNBC has the best configuration to accurately capture the underlying drought dynamics within the data, while avoiding the threat of overfitting. The goal is to find the most appropriate level of model complexity, determined by the number of latent drought states S_t (denoted by m), and to select suitable hyperparameters governing the rolling window sizes of the SPI and SDI input indices. The overarching goal here is to strike a balance between model complexity and goodness of fit.

Latent-State Cardinality m

As discussed in Section 1.3, the output of each latent RV (S_t) can be interpreted along a conceptual scale that is binned into categories. This means that larger values of m allows for finer grained bins and thus gives the model a greater ability to capture subtle drought dynamics. However, this increases the number of free parameters and thus the risk of overfitting. Conversely, a smaller m enforces simplicity but may fail to capture the full complexity of the observed data.

To guide this choice, three widely used criteria were applied: the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the maximised log-likelihood of the fitted model [55]. The equations for AIC and BIC are given by:

$$AIC = -2 \cdot \log L(\Theta) + 2p, \tag{3.12}$$

$$BIC = -2 \cdot \log L(\Theta) + p \cdot \log k, \tag{3.13}$$

where $L(\Theta)$ is the maximised value of the likelihood function, p is the number of free parameters, and k is the number of data points.

Both AIC and BIC embody the principle of Occam's razor [44], which favors models that achieve high likelihood with minimal complexity. BIC applies a stronger penalty on complexity and is thus generally favoured for conservative model selection. This study followed the following framework for selecting m:

1. **Primary:** Choose 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 improvements in $\log L(\Theta)$ diminish as m increases, the simpler model is preferred (this is the well known "elbow rule").

Rolling-Window Hyperparameters

The rolling windows applied to the SPI and SDI indices introduce additional hyperparameters for the problem. Each index can be computed over 3-, 6-, 9-, or 12-month windows, and any combination of these can be paired to form the input configuration for the DNBC. Unlike m, however, varying the rolling window lengths does not alter the number of free parameters in the model. Thus, criteria such as AIC or BIC are not meaningful here, and only the maximised log-likelihood of the fitted model is used to compare configurations.

Empirically, these two types of hyperparameters, m and rolling-window size combination, are decoupled. The value of m directly controls the capacity of the model, while the window sizes control the temporal aggregation of the input data. Since one affects the input domain and the other the latent structure, their influence on the overall likelihood are independent. In practice, this means the choice for m and the window size combination can be selected in isolation of each other, the former using AIC, BIC and log-likelihoods and the latter using only log-likelihoods. While a quantitative proof of full independence is non-trivial, it is not discussed here. Therefore, the model selection relies solely on this conceptual argument of decoupling.

Determining k and p

The quantities k and p in Equations (3.12)–(3.13) are defined following common conventions in the literature and implementation libraries such as `seqHMM` [56, 57].

The number of data points k is taken as the total number of number of data points and is calculated as $k = T \times 3$.

The number of free parameters p corresponds to the model's degrees of freedom. In the DNBC, this includes contributions from the prior, transition, and emission probabilities:

$$\begin{aligned} p &= (m-1) + m(m-1) + \sum_{n=1}^{3} m(C_n - 1) \\ &= m^2 - 1 + m \sum_{n=1}^{3} (C_n - 1), \end{aligned}$$

where C_n is the cardinality, or number of categories, of the n-th input index.

Log-Likelihood Estimation

The log-likelihood, $\ell(\Theta)$, measures the probability of the observed data given the current model parameters:

$$\ell(\Theta) = \log p(\mathbf{A}_{1:T} \mid \Theta).$$

Obtaining the maximised log-likelihood $L(\Theta)$ from this will be discussed later in Section 3.3.2.

The value of $\ell(\Theta)$ can be obtained directly from the messages computed during inference using the JT algorithm. After completing the message-passing procedure, a single downward message, $\delta_{\downarrow T}(S_T)$, is sent to fully calibrate the final cluster $\psi_T(S_T, \mathbf{A}_T)$ (it may be useful to see the JT in Figure 3.3). This ensures that $\psi_T(S_T, \mathbf{A}_T)$ represents the true posterior incorporating all evidence in the model. Marginalising over the final latent RV S_T yields:

$$p(\mathbf{A}_{1:T} \mid \Theta) = \sum_{S_T} \psi_T(S_T, \mathbf{A}_T),$$

from which $\ell(\Theta) = \log p(\mathbf{A}_{1:T} \mid \Theta)$ follows directly.

3.2.5. Model Output

The final point of discussion for this section revolves around the model outputs which were exported in two forms:

- Posterior decoding using the maximum posterior marginal (MPM) rule.
- State sequence decoding using the Viterbi algorithm.

1. MPM rule The MPM rule involves computing the point-wise marginal for the latent drought states S_t at each time step:

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

which is obtained using the calibrated JT outlined in Section 3.2.2 [38, 55].

Conceptually, this rule will pick the most likely state, with a confidence attached, at each time step independently. This is particularly useful in quantifying the model's uncertainty when making particular classifications. It should be noted that this rule often leads to an unlikely or even impossible state sequence. For example, say $S_t = m$ (which signals very wet conditions) is then followed by $S_{t+1} = 1$ (which signals very dry conditions) in the very next time step.

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:

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

This decoding provides a more realistic representation of drought progression [58].

3.3. Implementation

Having established the structure, inference mechanisms, and parameter estimation procedures of the proposed model, this section describes its practical implementation which translates the formulations above into code. The discussion that follows focuses on the key design decisions and computational procedures rather than the underlying mathematics, which have already been presented in Section 3.2.

3.3.1. Programming Environment and 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, developed by Johan du Preez and Corné van Daalen, is still under development but provides robust functionality for probabilistic graphical models. The `C++` implementation handled the construction of factors, JT message passing, parameter estimation, model selection, and extraction of posterior outputs in both forms described in Section 3.2.5.

Python was used to complement this workflow, particularly for data-related tasks including raw data extraction, preprocessing into model inputs, postprocessing of model outputs, and visualisation of results. The primary Python libraries used were `numpy` for numerical computations, `pandas` for structured data handling, transformation, and aggregation, and `matplotlib` for generating all figures and visualisations.

This division allowed `C++` to focus on core model computation while Python streamlined data management and analysis for the rest of the project.

3.3.2. Model Implementation

The model implementation consisted of two main components: the EM algorithm for parameter estimation, and model selection procedures for determining the optimal number of hidden drought states (m) and rolling-window size combination for the input indices SPI and SDI.

EM Algorithm

The EM algorithm was implemented as specified in Section 3.2.3, with the following decisions made:

- **Initialisation:** Model parameters were randomly initialised from a standard Gaussian distribution following common practice [46].
- **Convergence criteria:** The algorithm was terminated when the relative change in log-likelihood satisfied:
$$\frac{|\ell(\Theta)^{\text{new}} - \ell(\Theta)^{\text{old}}|}{|\ell(\Theta)^{\text{old}}|} < 10^{-4},$$
or when a maximum of 100 iterations was reached.

The corresponding pseudocode (Algorithm 4.1) summarises this process and is included for completeness.

Model Selection

Two independent model selection procedures were conducted. The first procedure involved sweeping the number of hidden states, m, from 2 to 10. For each competing model, the AIC, BIC, and maximum log-likelihood were computed. The second procedure evaluated combinations of rolling-window sizes for the SPI and SDI indices, each allowed to take values of 3-, 6-, 9-, and 12-month windows. Competing models were compared solely using the maximum log-likelihood criterion.

As discussed in Section 3.2.4, these two groups of hyperparameters are decoupled meaning that varying one has no effect on the other. Accordingly, the first procedure used a standard 3-month rolling window for both SPI and SDI indices, while the second procedure was done with the optimal value for m found in the first. The general framework below was applied to both procedures:

1. Each candidate configuration was run with 10 random restarts to mitigate local optima.
2. The run yielding the highest log-likelihood was retained and used as the maximised log-likelihood, $L(\Theta)$.
3. This maximised value was then used to compute AIC and BIC (where applicable) and to rank competing configurations.

Chapter 4

Results

This chapter presents the results obtained from the dynamic naive Bayes classifier (DNBC). The analysis begins with the outputs from model selection then moves to qualitative and quantitative assessment of the model's classification performance against the input drought indices used as a benchmark.

4.1. Model Selection and State Definition

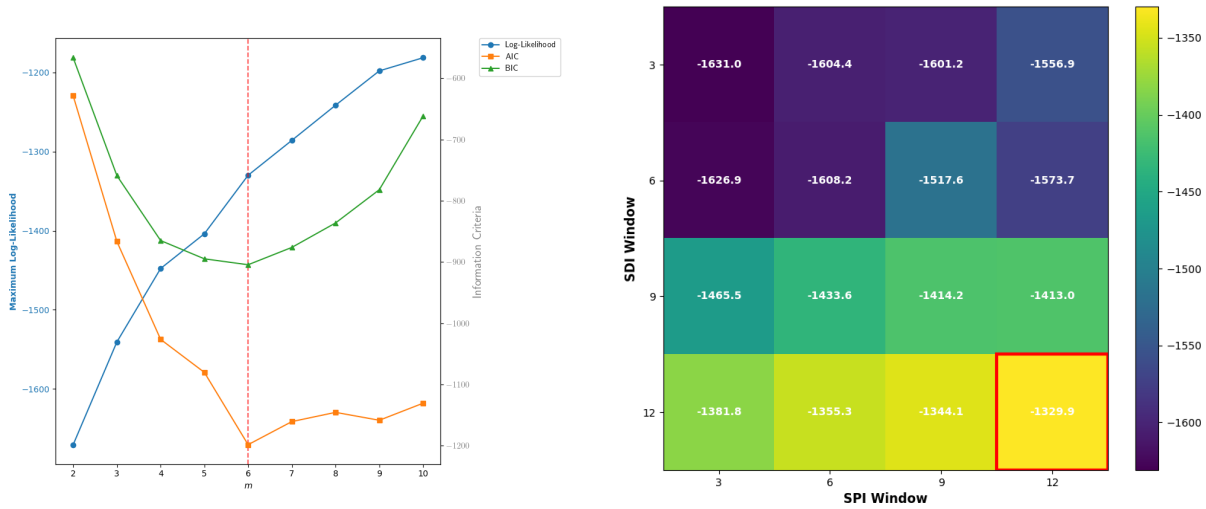
The results of the model selection procedures from Section 3.3.2 are presented in Figure 4.1. The plot on the left was used to determine the number of latent drought states, m. Both Akaike information criterion (AIC) and Bayesian information criterion (BIC) reach a minimum at $m = 6$, indicated by the vertical red line. This point also corresponds to a pronounced improvement in the log-likelihood. This alignment of all three selection criteria—following the rules established in Section 3.2.4—confirms that $m = 6$ is the optimal number of latent states.

The right panel shows the maximum log-likelihood for different combinations of SPI and SDI rolling-window sizes (3-, 6-, 9-, and 12-month), visualised as a heatmap. The results indicate that the 12-month window for both indices yields the highest likelihood, highlighted by the red box.

With the optimal number of latent drought states determined as $m = 6$, subsequent analyses use the following classification of latent states:

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

To be explicit, for the performance analysis that follows, latent states 1–3 ($S_t \leq 3$) were classified as "drought", while states 4–6 ($S_t > 3$) were classified as "non-drought", as



(a) Model selection criteria (AIC, BIC, and maximum log-likelihood) plotted against the number of hidden states m. The optimal value of m is indicated by a vertical red dotted line.

(b) Maximum log-likelihood values for different SPI and SDI window combinations. The 12-month window for both indices (red box) yields the optimal model performance.

Figure 4.1: Model selection results comparing different hyperparameters.

mentioned in Section 1.3.

4.2. Model Behaviour

This section qualitatively examines the performance and interpretability of the DNBC by looking at how it behaves when exposed to the data. Specifically, the Viterbi-decoded state sequence and the relationship between the model's output and the original input indices is explored.

4.2.1. Latent-State Sequence Output

Figure 4.2 presents the Viterbi-decoded state sequence for the entire study period 1981 to 2019, along with the known historical drought periods identified in literature as shaded regions, that being 1983–1984, 1991–1992, 1994–1995, 2000–2001, 2003–2004, 2014–2016, and 2017–2018.

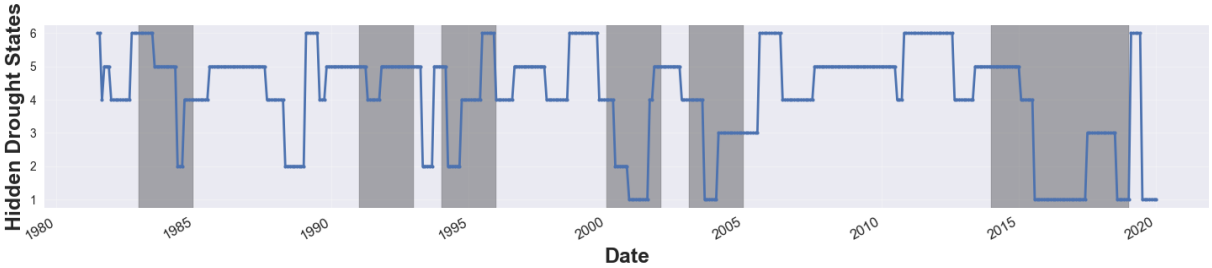


Figure 4.2: Viterbi-decoded state sequence of the DNBC with known drought periods as shaded regions.

The model demonstrates a mixed but promising capacity for identifying historical drought periods. It successfully captures the two major droughts in the 2000–2004 period, with latent states distinctly shifting lower. Performance for other events is more varied as the model identifies the 1983–1984 and 2014–2018 droughts, albeit with a noticeable delay, and fails to react to the 1991–1992 event completely.

Despite these inconsistencies, the model exhibits a clear tendency to enter lower drought states during known drought events. This indicates that the DNBC has learned to characterise drought conditions to some extent, successfully integrating the three input indices.

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.

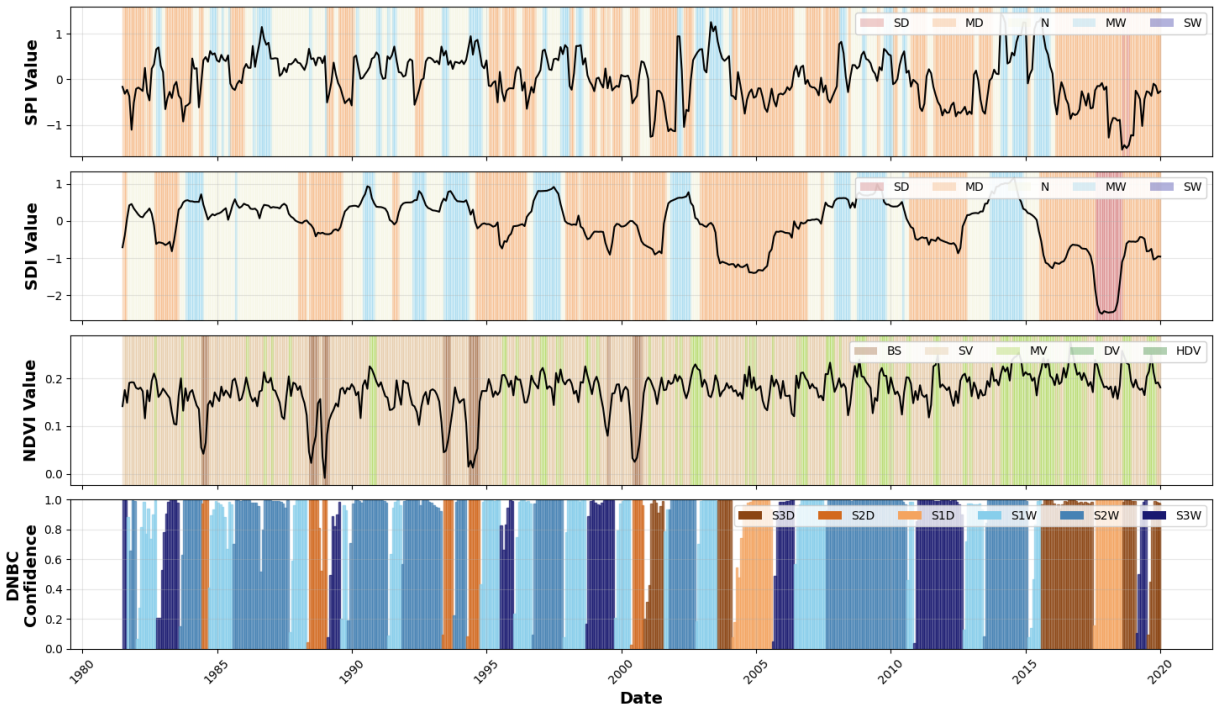


Figure 4.3: Drought classifications for the period 1981–2019 using SPI, SDI, NDVI, and DNBC (the vertical, coloured bars represent different drought states, while their height indicates the confidence in classification. The black lines plot the continuous values of SPI, SDI and NDVI)

This graph shows that the input indices themselves are inherently noisy and oscillatory, reflecting the high variability of environmental conditions and could in turn contribute to the model's uncertain classifications that are present in the plot, particularly at state

transitions. Nonetheless, this probabilistic approach allows these uncertainties to be explicitly shown, which the 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 the 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.

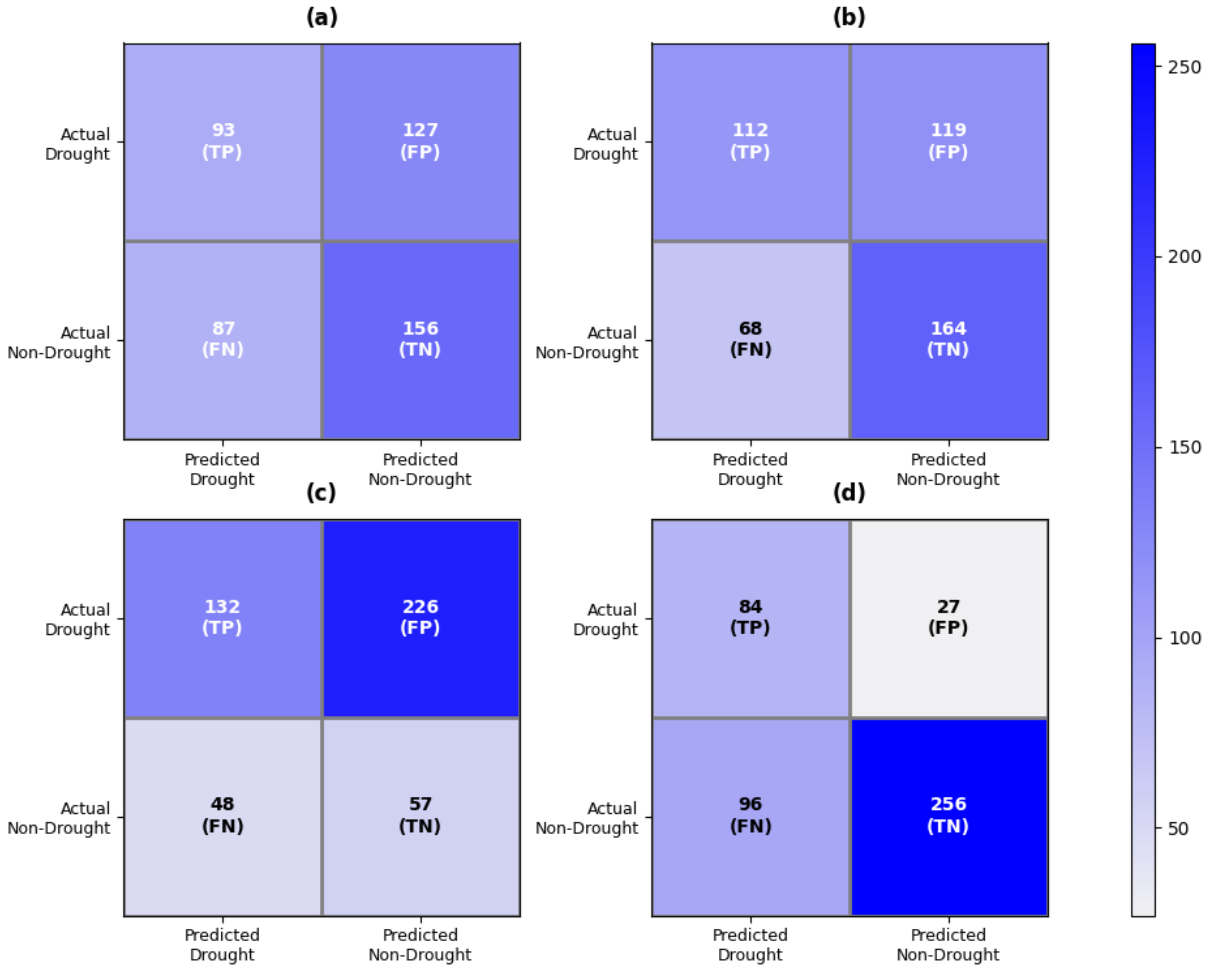


Figure 4.4: Confusion matrices for classifying known drought states using (a) SPI, (b) SDI, (c) NDVI, (d) DNBC.

4.3.1. Performance Metrics and Interpretation

Recall and precision evaluate two distinct and critical aspects of model performance in the context of drought monitoring. Recall measures the model's ability to correctly identify

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	51.67	53.78	42.27	46.50
SDI	62.22	59.61	48.48	54.50
NDVI	73.33	40.82	36.87	49.07
DNBC	46.67	73.43	75.68	57.73

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 for decision-makers.

The F1-score, defined as the harmonic mean of precision and recall, provides a single metric that balances these two concerns. It is particularly informative in highly imbalanced datasets such as this one where drought events form a small fraction of the total. Finally, 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 (73.3%) but low precision (36.9%), 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 moderate recall (51.7%) paired with lower precision (42.3%), suggesting the same behaviour as the NDVI, just to a lesser extent.

- **SDI:** Achieves higher values than the SPI for both recall (62.2%) and precision (48.5%). SDI therefore offers a slight improvement to the SPI, as shown by the increased F1-score (54.5%)

- **DNBC Output (Viterbi):** Marginally achieves the highest F1-score (57.7%), with the lowest recall (46.67%) but the largest precision (75.7%). This suggests the model misses true drought events, but its drought classifications are reliable, resulting in a conservative classifier.

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 in the southwestern Cape region using open source datasets. Raw precipitation and streamflow data were collected, cleaned and formatted for index calculation. In contrast, the NDVI was obtained directly from a dataset provided by the National Oceanic and Atmospheric Administration (NOAA). These three indices were then discretised which together formed the input dataset for the DNBC model. The model was formulated with discrete random variables (RVs) representing latent drought states and observed input indices. Parameter estimation was performed using the expectation-maximisation (EM) algorithm paired with the junction tree (JT) algorithm for inference, and model selection was guided by Akaike information criterion (AIC), Bayesian information criterion (BIC), and maximised log-likelihood criteria.

The model was evaluated both qualitatively and quantitatively, with plots revealing that the DNBC successfully identified known drought periods in the Western Cape. The resulting DNBC was reliable when it predicted drought, but missed a number of true drought events. 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 successfully captured abstract information across the different drought dimensions.

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 RVs, for example via Gaussian or hybrid emission distributions, could increase the model's capacity to capture underlying drought dynamics.

- **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 dynamic naive Bayes classifier, represent a viable approach to drought characterisation in data-limited environments. While challenges remain, the results show that integrating multiple drought dimensions using a principled and probabilistic approach yields both interpretive and operational value. With further refinement 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] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[39] Z. Hao and A. AghaKouchak, "Multivariate standardized drought index: A parametric multi-index model," *Advances in Water Resources*, vol. 57, pp. 12–18, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0309170813000493

[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] 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

[42] 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.

[43] E. Vermote and N. C. Program, "Noaa climate data record (cdr) of avhrr normalized difference vegetation index (ndvi), version 5," 2019. [Online]. Available: https://doi.org/10.7289/V5ZG6QH9

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

[45] 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.

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

[47] 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

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

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

[50] 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.

[51] 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

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

[53] 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.

[54] 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

[55] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[56] 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.

[57] 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.

[58] 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

WBS Number	Task name / Title	Planned start date	Planned end date	Progress (%)	Duration (hrs)	Status	Predecessor
1	**Phase 1: Foundation and Problem Understa**	**2025/07/21**	**2025/08/03**	**100**	**126**		
1.1	Reading Literature Around Drought	2025/07/21	2025/07/25	100	45	Done	
1.2	Learning HMMs	2025/07/26	2025/07/31	100	54	Done	1.1
1.3	Learning DNBCs	2025/08/01	2025/08/03	100	27	Done	1.2
1.4	Learning Indice Calculations	2025/07/26	2025/07/27	100	18	Done	1.1
2	**Phase 2: Data Acquisition and Preprocessin**	**2025/08/04**	**2025/08/13**	**100**	**90**		**1**
2.1	Identifying and Sourcing Datasets	2025/08/04	2025/08/07	100	36	Done	
2.2	Data Scraping and Cleaning	2025/08/08	2025/08/11	100	36	Done	2.1
2.3	Exploratory Data Analysis	2025/08/12	2025/08/13	100	18	Done	2.2
3	**Phase 3: Index Computation**	**2025/08/14**	**2025/08/25**	**100**	**108**		**2**
3.1	SPI Calculation	2025/08/14	2025/08/16	100	27	Done	
3.2	SDI Calculation	2025/08/17	2025/08/18	100	18	Done	3.1
3.3	NDVI Extraction	2025/08/19	2025/08/23	100	45	Done	3.2
3.4	Discretisation	2025/08/24	2025/08/25	100	18	Done	3.3
4	**Phase 4: Model Implementation**	**2025/08/26**	**2025/09/17**	**100**	**207**		**3**
4.1	Model Formulation	2025/08/26	2025/08/31	100	54	Done	
4.2	Parameter Estimation	2025/09/01	2025/09/08	100	72	Done	4.1
4.3	Inference and Decoding	2025/09/09	2025/09/14	100	54	Done	4.2
4.4	Model Selection and Validation	2025/09/15	2025/09/17	100	27	Done	4.3
5	**Phase 5: Evaluation and Analysis**	**2025/09/18**	**2025/10/01**	**100**	**126**		**4**
5.1	Qualitative Evaluation	2025/09/18	2025/09/22	100	45	Done	
5.2	Quantitative Evaluation	2025/09/23	2025/09/25	100	27	Done	5.1
5.3	Results Analysis	2025/09/26	2025/10/01	100	54	Done	5.2
6	**Phase 6: Documentation and Reporting**	**2025/10/02**	**2025/10/29**	**100**	**252**		**5**
6.1	Drafting Report	2025/10/02	2025/10/11	100	90	Done	
6.2	Revisions and Finalisation	2025/10/12	2025/11/02	100	162	Done	6.1
7	**Phase 7: Submission**	**2025/10/30**	**2025/11/19**	**0**	**189**		
7.1	Submit Report	2025/11/03	2025/11/03	0	9	Open	6.2
7.2	Oral Presentation Preperation	2025/11/02	2025/11/07	0	72	Open	6.2
7.3	Poster Creation	2025/11/02	2025/11/08	0	36	Open	6.2
7.4	Submit Slides & Oral Presentation	2025/11/09	2025/11/09	0	9	Open	7.2
7.5	Submit Poster	2025/11/09	2025/11/09	0	9	Open	7.3
7.6	In Person Orals	2025/11/10	2025/11/14	0	36	Open	
7.7	Project Open Day	2025/11/19	2025/11/19	0	9	Open	

Figure A.1: Project Plan Schedule

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 3.2.1). 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 3.2.1 and Section 3.2.3). 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 3.1.4). 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 3.2.1). 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 3.2.4). 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, index computation and discretisation (Section 3.1), 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 3.3).

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 3.2). Additionally, significant effort was spent understanding drought indices (SPI, SDI, NDVI), their derivation, and interpretation within the South African context (Section 3.1.4). This demonstrates a high level of independent learning ability and adaptability to new technical challenges.

Appendix C

Project Design Diagram

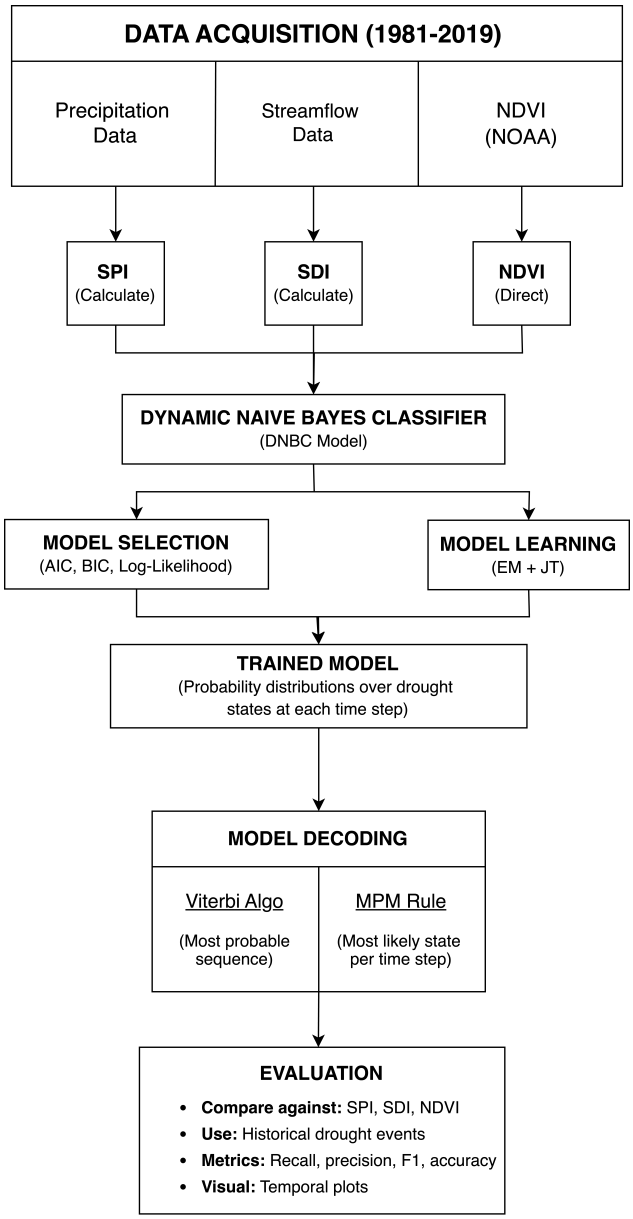


Figure C.1: End-to-end workflow of the project design. The pipeline encompasses data acquisition, drought index calculation (SPI and SDI), DNBC model training and selection, output decoding, and evaluation using standard classification metrics.

Appendix D

Expectation–Maximisation Algorithm for Model

Algorithm 4.1: Expectation–Maximisation (EM) Procedure for Dynamic Naive Bayes Classifier (DNBC)

Initialise parameters $\Theta^{\text{old}} \sim \mathcal{N}(0, 1)$
for $t = 1$ to 100 **do**
    Compute posterior distributions $p(S_t)$ and $p(S_t, S_{t+1})$ via Junction Tree inference keeping the parameters Θ^{old} constant
    While keeping posterior distributions constant, compute the updated parameters Θ^{new} using the re-estimation update rules
    Compute log-likelihoods $\ell(\Theta^{\text{old}})$ and $\ell(\Theta^{\text{new}})$
    if $\frac{|\ell(\Theta^{\text{new}}) - \ell(\Theta^{\text{old}})|}{|\ell(\Theta^{\text{old}})|} < 10^{-4}$ **then**
        break
    end if
    Update parameters $\Theta^{\text{old}} \leftarrow \Theta^{\text{new}}$
end for
return Θ^{new}, $\ell(\Theta^{\text{new}})$